AD_____
(Leave blank)

Award Number:
W81XWH-08-1-0428

TITLE:
Genetic and Environmental Pathways in Type 1 Diabetes Complications

PRINCIPAL INVESTIGATOR:
Massimo Trucco, M.D.

CONTRACTING ORGANIZATION:
University of Pittsburgh
Pittsburgh, PA 15260

REPORT DATE:
September 2010

TYPE OF REPORT:
Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: (Check one)

X Approved for public release; distribution unlimited

☐ Distribution limited to U.S. Government agencies only; report contains proprietary information

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01-09-2010	Final	1 SEP 2008 - 31 AUG 2010

4. TITLE AND SUBTITLE
Genetic and Environmental Pathways in Type 1 Diabetes Complications

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-08-1-0428

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Massimo Trucco, M.D.

Email: mnt@pitt.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Pittsburgh
Office of Research
350 Thackeray Hall
Pittsburgh, PA 15260

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research
and Materiel Command
Fort Detrick, MD 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Genetic factors contribute to risk for developing nephropathy in patients with Type 1 Diabetes (T1D). Cigarette smoking is deleterious to kidney function and is a risk factor for Diabetic-Nephropathy (DN) as well as End Stage Renal Disease (ESRD) in patients with T1D. The proposed study investigates how environmental exposure(s) (e.g., smoking) and genetic variants interact to amplify risk for T1DN and substantially increase incidence of ESRD. The specific aims are: 1) Identify genetic variants conferring risk to T1DN by performing a staged follow-up of our initial Genome-Wide Association Scan (GWAS) results; 2) Ensure that SNPs identified by Aim 1 affect risk of T1DN, as opposed to risk for T1D; 3) Identify genetic variants that interact with smoking status in conferring risk for T1DN; 4) Confirm results obtained during Aims 1-3 using an independent cohort of case and control participants. The relevance of the study to public health is that 16 million people in the US have diabetes with 800,000 new cases diagnosed each year. Diabetic complications threatening vision, kidney, and nerve function affect most diabetic patients. Improved prediction of risk for developing diabetes and diabetic complications among active duty members of the military, their families and retired military personnel will potentially allow focused preventative treatment of at-risk individuals, providing significant healthcare savings and improved patient well being.

15. SUBJECT TERMS
End Stage Renal Disease; Genetic Association; Genome Scanning; Nephropathy; Type 1 Diabetes

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
U	U	U	UU	96	**19b. TELEPHONE NUMBER** *(include area code)*

University of Pittsburgh
W81XWH-08-1-0428
Final Report (09/01/2008 – 08/31/2010)
Table of Contents

Year 01 (September 1, 2008 – August 30, 2009) ……..……...…	**4**
Year 02 (September 1, 2009 – August 30, 2010) ……..……...…	**46**
Bibliography of Publications ……..……..	**59**
List of Personnel……...	**60**
Reprints of Publications…….……	**61**

INTRODUCTION:

From our originally submitted Statement of Work:

To accomplish our Technical Objective 4 we will complete 4 additional tasks:

Task 6. Select reproducible SNPs and genes identified during Aims 1-3 of the previously funded proposal. Use biological information about the kidney along with genetic analyses of Stage 1 Genome-Wide Association Scan (GWAS) and Stage 2 Case/Control data to identify genomic regions with corresponding p-values approaching genome-wide significance.

Task 7. Identify confirmed genetic variants for T1DN. Using a cohort of case (N=150) and control (N=150) participants employ TaqMan methodology to genotype the SNPs reproducibly associated with T1DN. Statistical criteria for final decision will be a p-value *exceeding* genome wide significance for association. Replicate the genotype of select SNPs using an alternative genotyping method (i.e., Pyrosequencing).

Task 8. Identify confirmed haplotypes associated with T1DN. Perform comparative genome analysis to identify evolutionary conserved regions likely to regulate genetic penetrance of T1DN.

Task 9. Identify causal genetic variants for T1DN. Use GEnetic Matching (GEM) algorithms to identify a set of highly matched case and control participants. Initiate molecular characterization of confirmed genomic haplotypes by combined fine mapping of polymorphisms and sequencing of select evolutionary conserved regions of genomic DNA.

The research goals for the 3rd year funding are to use the cohort of T1D cases and controls as well as T1D case families to perform fine mapping of T1D association signals discovered during the genome-wide association (GWA) scan performed during years 1 and 2 of the project. The work performed during the recently completed research period was based upon Task 6 (Select reproducible SNPs and genes identified during the GWA scan), Task 7 (Identify confirmed genetic variants for the disease phenotype), and Task 8 (Identify haplotypes associated with Type 1 Diabetes and complications). Our progress towards this goal is addressed below.

BODY:

Our first quarterly scientific progress report (09/01/08 – 11/30/08) detailed the following steps forward in reaching the aims of our study.

Goal 1. Finalize the recruitment of T1D affected singletons and family trios by the middle of the 2009 research period. **Milestone 1A.** During the first 2 research quarters of the third year we will continue to pursue the recruitment of new T1D case trios and singletons from Pittsburgh. **Milestone 1B.** During the next research period we will continue to arrange new collaborations with researchers at other institutions (e.g., Wellcome Trust Case Control Consortium and EMIL researchers at Ulm, Germany) to gain increased access to DNA repositories and datasets useful for evaluating the genetics of T1D.

We now have the final tally for DNA samples available from our collaborators in Ulm and Frankfurt, Germany as well as Sicily, Italy. These samples represent the complete cohort of cases and controls that will be used for replication and fine mapping studies of the genetic signals identified during years 1 and 2 of the project. As summarized in Table 1 there are a total of N=3,942 T1D cases, N=3,750 non-T1D controls, and N=869 T1D case family trios that are available. The completed cohort includes the cohort of N=1,174 T1D case singletons and N=569 T1D case family trios recruited previously in the U.S. during the first 2 years of the project. Of these samples the entire U.S. cohort is currently available. In contrast, N=768 T1D singleton cases and N=1,750 singleton controls, representing 56% of the samples, from Ulm, Germany are currently available while the remaining samples from Germany and Sicily are scheduled for shipment during the Spring of this year.

Table 1. Summary of DNA samples available for the T1D project[1].

	Singletons		Family Trios
Population	Case	Control	Case
Frankfurt, Germany	500	500	----
Sicily, Italy	500	500	----
Ulm, Germany	1,768	2,750	300
U.S.	1,174	----	569
Total Samples	3,942	3,750	869

1. Index cases from singleton and family trios are listed as independent populations.

The data shown in Figure 1 summarizes the results of power analyses for estimating the likelihood that a genetic risk element (i.e., a SNP that is causal for T1D susceptibility) will be observed with 90% power and a p-value of less than 10^{-8}, the threshold for statistical significance in a GWA study. The x- and y-axes indicate the correlation between the relative risk of disease associated with a casual SNP and the total number of samples available to the study. The different curves indicate the threshold for 90% power when the minor allele frequency (MAF) of the causal SNP varies between 30% and 0.3%. As expected more samples are needed to observe a statistically significant signal from a SNP with small relative risk. Likewise, more samples are required when the causal SNP is less frequent (i.e., the minor allele exhibits a relatively smaller MAF value). As illustrated in Figure 1, the available cohort of N=8,561 samples is sufficient to provide 90% power to detect a causal SNP with an effective genetic risk ratio of 1.3 when the minor allele frequency is 30%, a risk ratio of 1.5 when the minor allele frequency is 10%, and a risk ratio of 1.8 when the minor allele frequency is 3%.

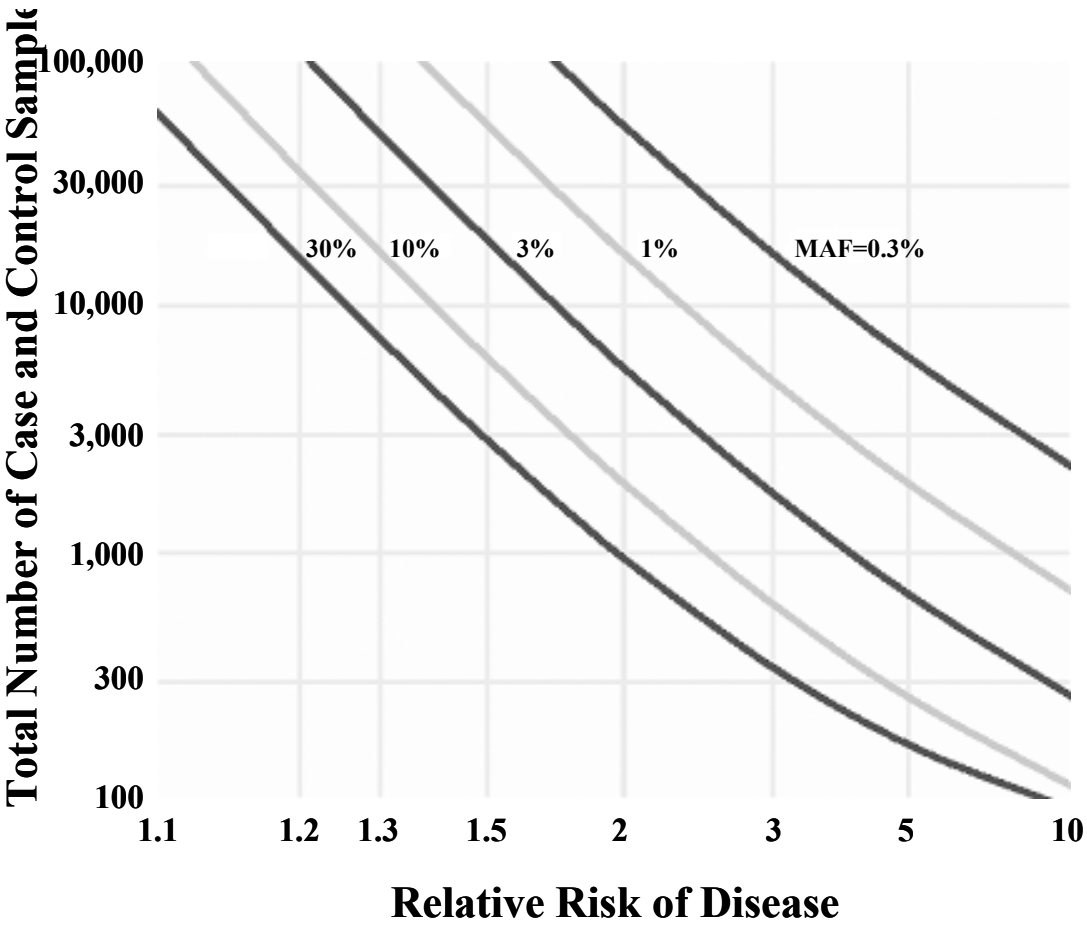


Figure 1. Sample size required for genetic association studies. The x-axis is the relative risk of disease associated with the causal SNP while the y-axis is the total number of samples available to the study. The curves represent the correlation between relative risk and sample number to achieve 90% power to detect a true signal with a p-value for association of $p<10^{-8}$. The abbreviation MAF is minor allele frequency. The figure is adapted from Altshuler et al. (2008).

In order to initiate fine mapping of T1D association signals we have compared the results from a variety of GWA studies that have been published over the past 3 years (Cooper et al., 2008; Todd et al., 2007; WTCCC, 2007). A summary of the p-values associated with different chromosomal regions is shown in Figure 2. These signals occur at a variety of positions along the human genome. The candidate loci for the strongest signals along with the corresponding SNP are indicate for 10 loci. Other strong signals that while not reaching genome-wide significance are also indicated (e.g., SNPs near the Insulin, *INS* and Interleukin-2 receptor A, *IL2RA* loci). In total there are 36 SNPs that have been assigned to mark 28 candidate loci for T1D susceptibility (Table 2).

The catalog of published T1D association signals is summarized in Table 2. Association between single nucleotide polymorphism and the T1D phenotype have been observed at 28 independent loci with the influence of genetic risk ranging between 1.07 for rs2165738 on Chromosome 2p23.3 to 8.3 associated with the HLA class II region on Chromosome 6p21.23. However, the mean relative risk factor for non-HLA loci is 1.3±2. Indicating that sample sizes of at least that available to the current cohort (listed in Table 1) and available to the third year research period will be required to effectively evaluate these genetic elements.

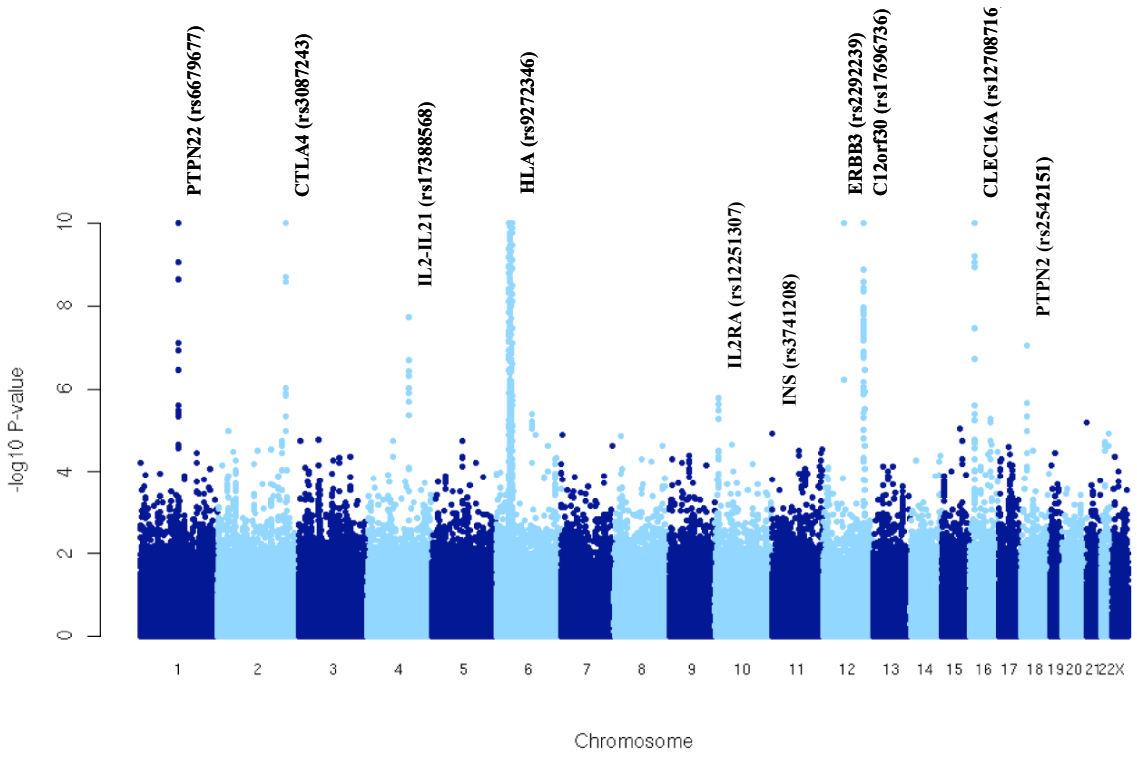


Figure 2. Results of GWA scan for SNPs associated with T1D. The results were adapted from the publication by Cooper et al. (2008) in which 6,225 cases and 6,946 controls were evaluated. The identity of candidate genes and SNPs associated with the strongest signals are indicated.

Table 2. Catalog of T1D association Signals.

dbSNP_ID	Chr	Location	p-Value	OR [95%CI]	Reported Genes(s)
rs1983853	1	85,083,780	2.0E-06	1.2 [1.11-1.29]	*LPAR3*
rs6679677	1	114,105,331	5.0E-26	1.82 [1.59-2.09]	*PTPN22*
rs2476601	1	114,179,091	1.0E-07	1.8 [1.44-2.24]	*PTPN22*
rs2165738	2	24,546,313	4.0E-06	1.07 [1.01-1.13]	*ITSN2, NCOA1*
rs9653442	2	100,191,799	5.0E-06	1.11 [1.05-1.17]	*AFF3, LOC150577*
rs1990760	2	162,832,297	2.0E-11	1.18 [1.11-1.23]	*IFIH1*
rs3087243	2	204,447,164	8.0E-11	Not Reported	*CTLA4*
rs6534347	4	123,417,885	2.0E-06	1.3 [1.10-1.55]	*IL2, IL21*
rs17388568	4	123,548,812	3.0E-06	1.26 [1.11-1.42]	*IL2, IL21*
rs6897932	5	35,910,332	8.0E-06	1.12 [1.06-1.19]	*IL7R*

rs1445898	5	35,946,286	8.0E-06	1.12 [1.06-1.19]	*CAPSL*
rs17166496	5	132,656,783	5.0E-06	1.3 [1.15-1.47]	*HSPA4, FSTL4*
rs9272346	6	32,679,809	0.0E+00	5.49 [4.83-6.24]	*HLA Class II*
rs2647044	6	32,738,427	1.0E-16	8.3 [6.97-9.89]	*HLA Class II*
rs3757247	6	91,014,184	1.0E-06	1.13 [1.08-1.19]	*BACH2*
rs11755527	6	91,014,952	5.0E-12	1.13 [1.08-1.19]	*BACH2*
rs10758593	9	4,282,083	3.0E-06	1.13 [1.07-1.19]	*GLIS3*
rs12251307	10	6,163,501	2.0E-06	Not Reported	*IL2RA*
rs947474	10	6,430,456	4.0E-09	1.1 [1.03-1.18]	*PRKCQ*
rs3741208	11	2,126,350	2.0E-07	1.25 [1.15-1.35]	*INS*
rs1004446	11	2,126,719	4.0E-09	1.61 [1.37-1.89]	*INS*
rs3764021	12	9,724,895	5.0E-08	1.57 [1.38-1.79]	*KLRB1, CELC2D, CLECL1, CD69*
rs11052552	12	9,747,225	7.0E-07	1.49 [1.28-1.73]	*KLRB1, CELC2D, CLECL1, CD69*
rs1701704	12	54,698,754	9.0E-10	1.25 [1.12-1.40]	*ERBB3*
rs11171739	12	54,756,892	1.0E-11	1.34 [1.17-1.54]	*ERBB3*
rs2292239	12	54,768,447	2.0E-20	1.28 [1.21-1.35]	*ERBB3*
rs17696736	12	110,971,201	2.0E-16	1.22 [1.15-1.28]	*C12orf30*
rs8035957	15	36,625,556	4.0E-06	1.14 [1.08-1.21]	*RASGRP1*
rs3825932	15	77,022,501	3.0E-15	1.16 [1.10-1.22]	*CTSH*
rs12708716	16	11,087,374	3.0E-18	1.23 [1.16-1.30]	*CLEC16A*
rs2903692	16	11,146,284	7.0E-11	1.54 [1.32-1.79]	*CLEC16A*
rs416603	16	11,271,580	3.0E-06	1.06 [1.01-1.12]	*TNP2, PRM3, C16orf75*
rs2542151	18	12,769,947	1.0E-14	1.3 [1.22-1.40]	*PTPN2*
rs763361	18	65,682,622	1.0E-08	1.16 [1.10-1.22]	*CD226*
rs9976767	21	42,709,459	2.0E-08	1.16 [1.10-1.22]	*UBASH3A*
rs229541	22	35,921,264	2.0E-08	1.04 [0.97-1.12]	*C1QTNF6*

Catalog of published T1D association signals reported by http://genome.gov (Revised Nov 25, 2008).

The critical goal for the next research period is to choose the regions of confirmed association for T1D susceptibility that will be studied using the cohort available for the third year research period. The criteria for this step are: firstly to identify regions in which SNPs listed in Table 2 may identify more than one candidate locus; and secondly to generate a list of tag-SNPs that can be used to fine map the chromosomal region(s) in order to define the most likely locus for T1D risk. For example, signals associated with Chromosome 4 (rs6534347 and rs17388568) have been associated with strong risk for disease (OR=1.3) but the reported data have been equivocal for whether Interleukin-2, *IL2* or Interleukin-21, *IL21* is more likely to represent the risk locus. A similar situation occurs on Chromosome 12 for rs3764021 (OR=1.6) and rs11052552 (OR=1.5) that have been assigned to at least 4 loci residing in this region of the genome.

In another example, the T1D association signals at rs11171739 and rs2292239 on Chromosome 12 have been identified as associated with the *ERBB3* locus. However, detailed analysis of the linkage disequilibrium (LD) structure of this chromosomal region indicated SNPs with strong LD (r^2>0.8) equally identified at least 2 nearby loci *RPS26* and *IKZF4*. The LD data for this region of the genome are summarized in Table 3 for rs11171739 and Table 4 for rs2292239. Moreover, SNPs with modest but significant LD in which r^2 is greater than 0.3 but less than 0.8 cover a broader range of genes and include *RAB5B* and *SUOX* in addition to the 3 loci mentioned previously. Because the ability to identify the most likely candidate allele is limited to the region tagged by the local genomic LD structure additional fine resolution mapping of genetic association data is required to unequivocally identify single genes. At this point in the project we will begin fine mapping of the regions identified in Table 3 and 4 in order to resolve the cluster of SNPs associated with T1D disease risk and to provide the data necessary to narrow down the casual region for disease susceptibility.

Table 3. LD Structure Surrounding rs11171739 Associated with T1D on Chromosome 12q13.2.

Location	dbSNP_ID	r^2	Distance from rs11171739	*Locus*	MAF
54,654,345	rs11171710	0.457	-102,547	*RAB5B*	0.458
54,655,773	rs773107	0.571	-101,119	*RAB5B*	0.275
54,656,178	rs773108	0.566	-100,714	*RAB5B*	0.276
54,660,962	rs773109	0.538	-95,930	*RAB5B*	0.283
54,665,327	rs773114	0.765	-91,565	*RAB5B*	0.367
54,665,694	rs1873914	0.765	-91,198	*RAB5B*	0.367
54,670,954	rs705698	0.480	-85,938	*RAB5B*	0.272
54,671,071	rs705699	0.790	-85,821	*RAB5B*	0.371
54,676,903	rs705702	0.538	-79,989	*SUOX*	0.283
54,687,352	**rs10876864**	0.862	-69,540	*Intergenic*	0.358
54,689,844	rs772921	0.624	-67,048	*Intergenic*	0.292
54,698,754	rs1701704	0.624	-58,138	*Intergenic*	0.292
54,703,195	rs2456973	0.621	-53,697	*IKZF4*	0.297
54,721,679	rs705704	0.613	-35,213	*RPS26*	0.280
54,722,196	**rs1131017**	0.927	-34,696	*RPS26*	0.377
54,753,854	rs7312770	0.699	-3,038	*Intergenic*	0.442
54,763,961	**rs2271194**	1.000	7,069	*ERBB3*	0.375
54,766,850	rs877636	0.704	9,958	*ERBB3*	0.293
54,768,447	rs2292239	0.714	11,555	*ERBB3*	0.300

SNPs located within 150kb of rs11171739 were evaluated for LD. Only those SNPs with an r^2 >0.3 are listed. SNPs with r^2>0.8 are indicated in bold font.

Table 4. LD Structure Surrounding rs2292239 Associated with T1D of Chromosome 12q13.2.

Location	dbSNP_ID	r^2	Distance from rs2292239	Locus	MAF
54,654,345	rs11171710	0.317	-114,102	*RAB5B*	0.458
54,655,773	**rs773107**	0.810	-112,674	*RAB5B*	0.275
54,656,178	rs773108	0.762	-112,269	*RAB5B*	0.276
54,660,962	rs773109	0.773	-107,485	*RAB5B*	0.283
54,665,327	rs773114	0.499	-103,120	*RAB5B*	0.367
54,665,694	rs1873914	0.499	-102,753	*RAB5B*	0.367
54,670,954	rs705698	0.715	-97,493	*RAB5B*	0.272
54,671,071	rs705699	0.518	-97,376	*RAB5B*	0.371
54,676,903	rs705702	0.773	-91,544	*SUOX*	0.283
54,687,352	rs10876864	0.577	-81,095	*Intergenic*	0.358
54,689,844	**rs772921**	0.884	-78,603	*Intergenic*	0.292
54,698,754	**rs1701704**	0.884	-69,693	*Intergenic*	0.292
54,703,195	**rs2456973**	0.883	-65,252	*IKZF4*	0.297
54,721,679	**rs705704**	0.879	-46,768	*RPS26*	0.280
54,722,196	rs1131017	0.636	-46,251	*RPS26*	0.377
54,753,854	rs7312770	0.488	-14,593	*Intergenic*	0.442
54,756,892	rs11171739	0.714	-11,555	*Intergenic*	0.375
54,763,961	rs2271194	0.714	-4,486	*ERBB3*	0.375
54,766,850	**rs877636**	1.000	-1,597	*ERBB3*	0.293
54,775,180	rs705708	0.388	6,733	*ERBB3*	0.475
54,794,676	rs4759228	0.421	26,229	*ERBB3*	0.242

SNPs located within 150kb of rs2292239 were evaluated for LD. Only those SNPs with an r^2 >0.3 are listed. SNPs with r^2>0.8 are indicated in bold font.

REFERENCES

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881-888.

Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, Allen JE, Downes K, Barrett JC, Healy BC, Mychaleckyj JC, Warram JH, Todd JA (2008) Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. Nat Genet 40:1399-1401.

Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, Lowe CE, Szeszko JS, Hafler JP, Zeitels L, Yang JH, Vella A, Nutland S, Stevens HE, Schuilenburg H, Coleman G, Maisuria M, Meadows W, Smink LJ, Healy B, Burren OS, Lam AA, Ovington NR, Allen J, Adlem E, Leung HT, Wallace C, Howson JM, Guja C, Ionescu-Tîrgovişte C; Genetics of Type 1 Diabetes in Finland, Simmonds MJ, Heward JM, Gough SC; Wellcome Trust Case Control Consortium, Dunger DB, Wicker LS, Clayton DG (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39:857-864.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678.

12. Use additional page (s) to present a brief statement of plans or milestones for the next quarter.

There are 2 goals that will be the focus of our work during the next research quarter.

Goal 1. Prepare available DNA samples for use in genotyping assays. **Milestone 1A.** Transfer DNA from screw capped stock tubes to barcode labeled freezer-tubes. **Milestone 1B.** Prepare working dilutions of DNA stocks compatible with 384-well analysis of SNPs. **Milestone 1C.** Transfer aliquots of DNA from tubes containing the working dilutions into 384-well genotyping reaction trays.

Goal 2. Prepare a set of SNPs for use in genetic fine mapping of T1D associated loci. **Milestone 2A.** Use the T1D cohort identified Table 1 to initiate genotyping of SNPs for fine mapping of T1D susceptibility signals.

In our second quarterly scientific progress report (12/01/08 – 02/28/09), we presented the following data:

There were 2 research goals for the preceding quarter: Goal 1) Prepare available DNA samples for use in genotyping assays; and Goal 2) Prepare a set of SNPs for use in genetic fine mapping of T1D associated loci. Our progress toward completing these tasks is addressed below.

Goal 1. *Prepare available DNA samples for use in genotyping assays.* ***Milestone 1A.*** *Transfer DNA from screw capped stock tubes to barcode labeled freezer-tubes.* ***Milestone 1B.*** *Prepare working dilutions of DNA stocks compatible with 384-well analysis of SNPs.* ***Milestone 1C.*** *Transfer aliquots of DNA from tubes containing the working dilutions into 384-well genotyping reaction trays.*

Completion of Milestones 1A and 1B: DNA samples were prepared from the complete collection of 6,219 samples (Table 1). The material has been transferred from the screw cap tubes, used for long term storage of the stocks, and into barcode labeled freezer tubes. The repository of DNA samples is stored in 65 storage boxes that hold 96 samples each. At the same time that the samples were transferred into barcode labeled tubes they were also diluted to a concentration of 2ng/ul that is suitable for use as a working dilution. The diluted samples are stored in a 4°C refrigerator until they are used for genotyping. These steps complete the Milestones listed as 1A and 1B that were presented in the preceding quarterly report.

Table 1. Formatted DNAs.

	Number
Total Samples	6,219
Storage Boxes	65
Concentration	2ng/μl
Amount	5μg

Continuation of Milestone 1C: Transfer of the aliquots of diluted DNA samples from tubes to the 384-well genotyping trays has not yet been performed. We decided to postpone this, the third milestone in Goal 1, until we had completed the work proposed as Goal 2. It is in Goal 2 that we will choose which SNPs will be subjected to intensive genotyping for

analysis of their association with T1D risk. Delay of Milestone 1C, however, presents only a small change to the overall project in that the work of transferring the samples from barcode tubes to reaction trays will be performed by the Beckman-Coulter BioMek FX liquid handling robot. This process will that a few days to a week to perform and will thus be accomplished quickly once it is time to initiate genotyping of the materials.

Goal 2. *Prepare a set of SNPs for use in genetic fine mapping of T1D associated loci.* ***Milestone 2A.*** *Use the T1D cohort identified Table 1 to initiate genotyping of SNPs for fine mapping of T1D susceptibility signals.*

Continuation of Milestone 2A: We have not yet completed the analysis of SNPs that will be used for fine mapping T1D risk elements. Therefore Milestone 2A will be continued during the next research quarter. We have, however, made substantial progress on selection of SNPs (Table 2). There are 3 SNPs rs4526299, rs224534, and rs2301151 located on Chromosomes 7, 17, and 19, respectively that will be used. These genetic markers have been identified through our collaborations with other T1D researchers and are anticipated to provide evidence for new risk loci.

Table 2. SNPs for fine mapping T1D risk elements.

dbSNP ID	Location	Locus	MAF
rs4526299	7:97,205,565	*TAC1*	0.223
rs224534	17:3,433,451	*TRPV1*	0.267
rs2301151	19:40,534,926	*FFAR1*	0.242

Complete analyses of the region marked by the SNPs listed in Table 2 are ongoing. For example the SNP rs4526299 on Chromosome 7 is in substantial linkage disequilibrium (LD) with 9 SNP, having an r^2 greater than 0.8 (Table 3) and is in statistically significant LD (r^2>0.3) with 35 SNPs covering a region of roughly 174kb and at least 2 loci (i.e., *TAC1* and *ASNS*).

Table 3. LD structure at rs4526299 (Chr7:97,205,565).

Location	dbSNP ID	r^2	Locus	MAF
97,167,342	rs1588422	0.584	*Intergenic*	0.317
97,187,959	**rs10260339**	0.854	*Intergenic*	0.246
97,194,047	rs13232655	0.665	*Intergenic*	0.321
97,203,867	**rs3779470**	1	*TAC1*	0.245
97,205,911	**rs4602814**	1	*TAC1*	0.237
97,206,685	**rs1510300**	0.873	*TAC1*	0.210
97,207,327	**rs12532490**	1	*TAC1*	0.231
97,207,432	**rs12534885**	1	*TAC1*	0.242
97,220,642	rs17169072	0.659	*Intergenic*	0.317
97,221,893	rs6465615	0.67	*Intergenic*	0.304
97,222,043	rs2894132	0.659	*Intergenic*	0.317
97,222,123	rs2394739	0.659	*Intergenic*	0.317
97,222,763	rs12536903	0.659	*Intergenic*	0.317
97,223,280	**rs10486008**	1	*Intergenic*	0.188
97,223,447	rs2017923	0.659	*Intergenic*	0.317
97,254,098	rs10245421	0.313	*intergenic*	0.491
97,261,330	rs1229540	0.338	*Intergenic*	0.491
97,272,564	rs6947443	0.316	*Intergenic*	0.500
97,275,564	rs2530156	0.321	*Intergenic*	0.491
97,290,463	**rs1510304**	0.904	*Intergenic*	0.267
97,297,477	rs714117	0.313	*Intergenic*	0.491
97,309,474	**rs7800746**	0.905	*Intergenic*	0.258
97,324,149	rs2074891	0.428	*ASNS*	0.417
97,325,282	rs2237289	0.444	*ASNS*	0.408
97,325,907	rs7792744	0.4	*ASNS*	0.400
97,326,669	rs7797354	0.385	*ASNS*	0.408
97,326,962	rs10253639	0.307	*ASNS*	0.392
97,330,582	rs7781469	0.383	*ASNS*	0.433

97,334,115	rs10263979	0.512	*ASNS*	0.342
97,337,018	rs3735557	0.371	*ASNS*	0.350
97,337,636	rs7790127	0.383	*ASNS*	0.433
97,338,010	rs11486966	0.38	*ASNS*	0.424
97,339,971	rs3757676	0.512	*ASNS*	0.342
97,341,076	rs3757674	0.512	*ASNS*	0.342
97,341,743	rs10234646	0.522	*Intergenic*	0.350

Similar to the preliminary analysis of the genomic region surrounding rs4526299, the regions surrounding the 2 remaining SNP (i.e., rs224534 on Chromosome 17 and rs2301151 on Chromosome 19) also tag a regions of the genome covering multiple loci and/or intergenic regions (Tables 4 and 5). For example, the SNP rs2301151 summarized in Table 5 covers a region of roughly 60kb and 2 loci (i.e., *CD22* and *FFAR1*). In contrast, the SNP identified for analysis on Chromosome 17 (Table 4) is in a region exhibiting smaller distances of LD and covers a single gene, TRPV1.

Table 4. LD structure at rs224534 (Chr17:3,433,451).

Location	dbSNP ID	r^2	Locus	MAF
3,431,117	rs150908	0.507	*TRPV1*	0.392
3,433,929	**rs224536**	0.958	*TRPV1*	0.246
3,433,951	**rs224537**	0.952	*TRPV1*	0.228
3,436,263	**rs161393**	1	*TRPV1*	0.254
3,439,747	**rs12936340**	0.868	*TRPV1*	0.259
3,439,949	rs222747	0.472	*TRPV1*	0.208

Table 5. LD structure at rs2301151 (Chr19:40,534,926).

Location	dbSNP ID	r^2	Locus	MAF
40,521,353	rs7251526	0.382	*CD22*	0.246
40,525,246	rs2312586	0.405	*CD22*	0.202
40,526,461	rs6510478	0.617	*CD22*	0.194
40,527,422	rs3746250	0.552	*CD22*	0.317
40,538,955	rs387083	0.592	*Intergenic*	0.350
40,556,175	rs453945	0.464	*FFAR1*	0.412
40,564,735	rs4806149	0.538	*Intergenic*	0.258
40,578,500	rs4806152	0.561	*Intergenic*	0.250
40,578,969	rs16970288	0.319	*Intergenic*	0.314
40,579,616	rs17304625	0.521	*Intergenic*	0.242
40,581,269	rs12978607	0.339	*Intergenic*	0.450

In order to complete Goal 2 of the recently concluded research quarter we will continue the computational analysis of the regions identified by the SNPs listed in Table 3 through 5 and will prioritize those SNPs for genotyping in order to efficiently tag these regions. Completion of this step will be accomplished shortly and will have the benefit of identifying a minimum number of genotyping assays that will be used to evaluate the association of these genomic regions with T1D risk.

12. Statement of Plans.

There are 3 goals that will be the focus of our work during the next research quarter.

Goal 1. Finalize SNPs for fine mapping T1D risk elements. **Milestone 1A.** Working with publicly available linkage disequilibrium data define SNP clusters and tag-SNPs for fine mapping of genomic regions associated with T1D.

Goal 2. Fine map genomic regions for association with T1D by genotyping select SNPs. **Milestone 2A.** Prepare DNA samples for laboratory analysis. **Milestone 2B.** Initiate genotyping of DNA samples.

Goal 3. Leverage the available data by preparing a new grant proposal to be submitted to the National Institutes of Health. **Milestone 3A.** Organize available preliminary data as well as the list of DNA and other biological materials available for the new research proposal. **Milestone 3B.** Prepare an appropriate hypothesis and set of specific aims to investigate the genetics of T1D. **Milestone 3C.** Prepare the research proposal and submit for the next grant deadline.

In our third quarterly scientific progress report (03/01/09 – 05/31/09) we then reported the following findings:

At the beginning of the recently completed research quarter we chose to modify the goals. The reason for this was to allow us to focus the majority of our effort on writing a grant application for the study of gene-gene interactions that influence Type 1 Diabetes (T1D) disease susceptibility. This was originally listed as Goal 3 for the research quarter but as a result of this modification became Goal 1. Moreover, a new second goal for the research quarter was added to the work effort. In that goal we designed an experiment for exploiting next generation pyrosequencing for analysis of genomic DNA regions that influence T1D risk. Both of these goals have been completed.

Of the 3 original goals that were proposed for the recently completed research quarter, one was kept (but revised for become Goal 1) and two were postponed to be included as research goals for the upcoming research period. The latter two goals can be found at this end of this quarterly report under the heading "Statement of Plans for the Upcoming Research Period".

Goal 1. *Leverage the available data by preparing a new grant proposal to be submitted to the National Institutes of Health.*

REFINING THE GENETIC AND FUNCTIONAL ARCHITECTURE OF TYPE 1 DIABETES
Principal Investigator: Massimo Trucco, MD

Goal 1. ABSTRACT
This application proposes to fine-map recognized Type 1 Diabetes (T1D) regions by using newly generated genetic and gene expression data as well as public data. It moves beyond fine-mapping in two ways: 1) we will further map expression quantitative trait loci (eQTL) relevant for T1D; and 2) using the eQTL information, the fine-mapped loci, and public resources we will uncover new T1D regions. Our investigative team has extensive experience in the study of T1D as well as in the development of statistical genetic methods required for fine-mapping. Team members have assembled an extensive cohort of T1D cases (N=10,457) and controls (N=6750). The available materials include DNAs, B-lymphoblastoid cell lines (B-LCLs), and peripheral blood mononuclear cells (PBMCs) that have already been prepared and are stored in the member laboratories. The cohort includes information on T1D status, gender, age of T1D onset, co-incident autoimmune disease(s), and environmental exposures (e.g., geographic location and date of T1D onset) together with HLA typing. Our study also creates an exciting resource for the proposed research and the wider community: we propose functional studies of gene expression and cellular responses using B-LCLs (acting as surrogates for antigen presenting cells) and autoreactive memory T-cells, both of which are relevant to T1D pathogenesis. The gene expression resources will be teamed with genome-wide association scan (GWAS) genotyping to aid in fine-mapping while enhancing efforts to discover new eQTLs linked with T1D etiology.

Goal 1. NARRATIVE
Type 1 Diabetes (T1D) is the second most common chronic disease of childhood. The project goal is to elucidate molecular networks affecting T1D susceptibility that are directly influenced by stably inherited genetic variants. Increased understanding of the gene-networks underlying disease risk will aid in the development of accurate screening tools as well as in creation of new therapeutic treatments.

Goal 1. PROJECT DESCRIPTION
This application proposes to fine-map recognized Type 1 Diabetes (T1D) regions by using newly generated genetic and gene expression data as well as public data. It moves beyond fine-mapping in two ways: 1) we will further map expression quantitative trait loci (eQTL) relevant for T1D; and 2) using the eQTL information, the fine-mapped loci, and public resources we will uncover new T1D regions. Our investigative team has extensive experience in the study of T1D (Trucco and Boehm) as well as in the development of statistical genetic methods required for fine-mapping (Devlin and Roeder). Team members have assembled an extensive cohort of T1D cases (N=10,457) and controls (N=6750). The available materials include DNAs, B-lymphoblastoid cell lines (B-LCLs), and peripheral blood mononuclear cells (PBMCs) that have already been prepared and are stored in the member laboratories. The cohort, a combined resource from team members Trucco and Boehm,

include information on T1D status, gender, age of T1D onset, co-incident autoimmune disease(s), and environmental exposures (e.g., geographic location and date of T1D onset) together with HLA typing. Comparable data have not been used for T1D fine-mapping, with the exception of Smyth et al.[1] Our study also creates an exciting resource for the proposed research and the wider community: we propose functional studies of gene expression and cellular responses using B-LCLs (acting as surrogates for antigen presenting cells) and autoreactive memory T-cells, both of which are relevant to T1D pathogenesis.[2] The gene expression resources will be teamed with genome-wide association scan (GWAS) genotyping to aid in fine-mapping while enhancing efforts to discover new eQTLs linked with T1D etiology.

Hypothesis: Causal pathways leading to T1D can be constructed by integration of data obtained through GWAS, genome-wide mRNA expression, as well as characterization of autoantigen presenting cells and autoreactive memory T-cell populations.

The goal is to elucidate molecular networks affecting T1D susceptibility that are directly influenced by stably inherited genetic variants (i.e., SNPs, small insertion/deletions, and copy number variants). DNA is available from T1D patient cases (N=7239), multiplex case families (N=1609), and non-T1D controls (N=6750) that have been recruited from populations in northwestern and southern Europe as well as the U.S. (Table 1). In addition to DNA, the available repository, summarized in Table 2, contains a substantial collection of B-LCLs (N=3768 case and N=300 control), and PBMCs (N=500 case and N=300 control). The assembled samples represent a unique cohort with sufficient power to: 1) fine-map previously recognized T1D risk loci; 2) identify T1D specific *cis* and *trans*-acting regulatory elements influencing eQTLs; 3) characterize the T1D specific functional properties of long lived autoreactive memory T-cells; and 4) exploit an integrative genomics approach to discover gene networks causal for T1D.

Table 1. Independent DNA samples.

	Singletons	Multiplex Families[1]
Population	Case/Control	Case
Frankfurt, Germany	500/500	
Ulm, Germany	1,768/2,750	1,300
Genoa, Italy		309
Sicily, Italy	500/500	
Pittsburgh, PA	4,471/3,000	

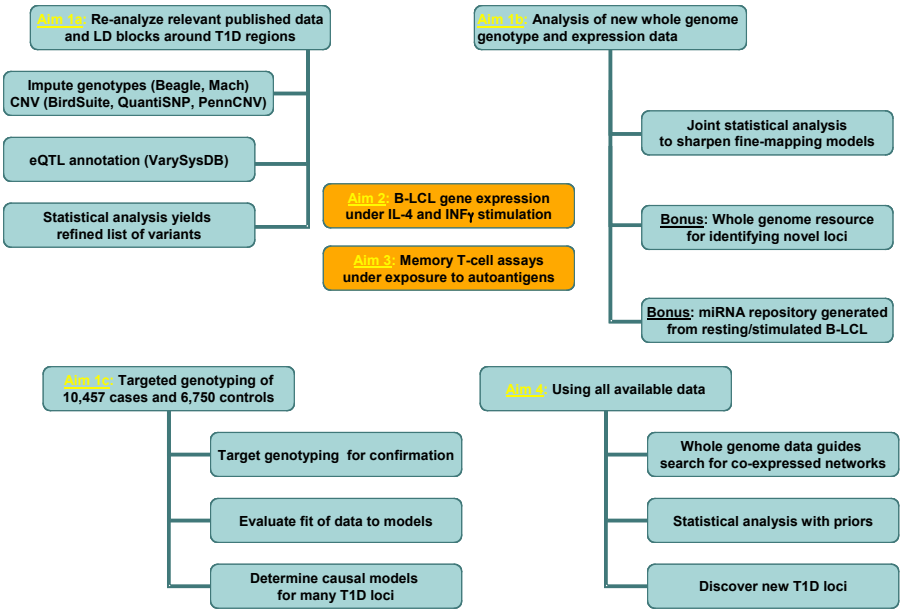
1. Multiplex families consist of 2 affected siblings and typically 2 parents.

Table 2. Biologic materials[1].

	Singletons	Multiplex Families
Material	Case/Control	Case
B-LCL	1,768/300	1,000
PBMC	500/300	----

1. Ulm, Germany.

The underlying assumption of the approach is that the response of T1D relevant cells (e.g., antigen presenting B-lymphocytes and autoreactive memory T-cells) will be reflected in disease specific perturbations of eQTLs and gene networks. The experimental design will focus on fine-mapping regions associated with T1D and will use the entire cohort of case and control DNA samples, roughly 17,207 samples (including a cohort of N=3218 affected siblings obtained from the collection of multiplex families), in addition to publicly available data from T1D studies (e.g., T1DGC, WTCCC and others).[3-8] In order to enable the discovery of eQTLs, a GWAS will be performed using N=600 representative participants chosen as described below. To fine map previously identified T1D loci, we will move beyond traditional eQTL data by characterizing cellular responses of B-cells and memory T-cells that are integrally involved in T1D pathogenesis under disease-appropriate conditions. To illuminate disease-specific gene networks critical for pathogenesis of T1D, we will integrate three kinds of approaches and data: 1) genotypes from case-control samples; 2) functional responses of mRNA and protein abundance (e.g., HLA class II abundance on antigen presenting cells, B-LCLs, and cytokine production by autoantigen stimulation of memory T-cells); and 3) covariate predictors of T1D (e.g., age and year of onset, geographic location, gender, and ancestry).



There are at least two approaches to fine-mapping: (A) dense sequencing of wide regions associated with risk for T1D, followed by further genetic, statistical and molecular analysis; or (B), statistical analysis of the wide regions using the publicly-available data to fine-map regions; followed by further fine-mapping in our "local" samples; then further genetic, statistical and molecular analysis. We wrestled with (A) versus (B) and decided that (B) is more consistent with the loci we seek. These loci were uncovered by GWAS powered to detect common variation affecting risk.[9] In the absence of very large sample sizes the power of GWAS studies to detect uncommon variants is poor, making it unlikely the loci we seek are uncommon. Most common variations are already known and any missed will be uncovered through the "1000 Genomes Project." Thus in our view little new will be discovered by additional, dense sequencing. **We will concentrate on evaluating known variations and measuring likely effects on risk and phenotypic/clinical characteristics for T1D.** An outline of how statistical models and biological data integrate to achieve our goals is illustrated in Figure 1.

T1D is a heterogeneous collection of disease processes with different clinical characteristics (e.g., age of onset and appearance of other autoimmune diseases) that delineate two major subgroups of patients.[10] Long-term follow up has revealed that about 1/3 of T1D patients recruited for our cohort have autoimmune thyroid disease. The genetic elements influencing autoimmune mediated destruction of tissues/organs involve shared risk elements beyond those described for the HLA region.[1,11] We can extend and improve the causal pathway analysis by subtype analysis of T1D populations that will include delineation of the common pathway leading to islet β cell autoimmunity and delineation of determinants that define age of onset because we have access to extensively phenotyped cohorts of patients and controls. We will address both loci and functional analysis with distinct effects on T1D only vs. T1D with other autoimmune diseases by investigating the model that common disease effector pathways occur between HLA-associated autoimmune disorders. We are uniquely positioned to contribute to fine-mapping T1D loci. We have deep experience in clinical presentation of T1D, its important external predictors, genetics, as well as the statistics and modeling of genetic data. We are a small team that can move nimbly to deliver results. We have contributed to the extensive methodological literature on fine-mapping, and we expect to do so again in this work, as well as fine-mapping T1D loci. *The Specific Aims of the project follow:*

Aim 1. Fine map T1D risk loci.
1a) Re-analyze all relevant and available GWAS data to refine signals for association and assess potential heterogeneity of association signals, such as by ancestry. Query eQTL databases for relevant associations.

Goals: **1a1**. Identify local regions harboring risk variants. **1a2**. Identify SNPs for large scale fine-mapping in new cohorts.

1b) Using the Affymetrix 6.0/Illumina 1M genotyping chip, genotype a selected set of 600 DNA samples from individuals whose cells were transformed into B-LCL and that have provided PBMC samples for analysis of memory T-cells. Combine these results with those of 1a as a means of confirming patterns of association.

Goals: **1b1**. Develop a resource for subsequent causal inference by the integration of genotypes produced on these samples with the gene expression and cellular functional data obtained in Aims 2 and 3. **1b2**. Garner additional support for genetic models inferred by analysis in 1a. **1b3**. Use quantitative phenotypes of gene

expression to further refine genetic models. *Additional benefit*: Perform GWAS on gene expression data to identify novel loci for T1D risk and contribute to the literature on the determinants of gene expression.

1c) Using targeted genotyping, confirm and/or refine models derived in 1a and 1b in new samples comprising roughly 10,457 cases and 6750 controls.

Goals: **1c**. Identify causal variants for evaluation in B-lymphoblastoid cell line systems.
Methodology used during Aim 1a of the project follow:

Definition of T1D loci: Linkage disequilibrium (LD) typically extends at most 50Kb in populations of European ancestry.[12,13] LD in all regions surrounding T1D-associated SNPs will be assessed using available databases. *Implementation:* our software, Hclust,[14] and Haploview.[15] We will take as the T1D locus the region containing all SNPs plausibly in LD with the T1D GWAS SNPs.

Quality Control: We plan consistent, aggressive QC on all data sets. We will eliminate genotypes/loci showing evidence for elevated genotyping error or high missing rates within a platform and correct for inconsistencies in genotype calling (e.g., different strand) across platforms. *Implementation*: our QC in-house software.

Table 3. Examples of public datasets.

Cohort	Case/Control
DCCT/EDIC[1]	1,411/----
GoKinD[2]	1,795/----
WTCCC[2]	1,963/2,938
Framingham SHARe[2]	----/1,529
KORA[2]	----/1,644
POPGEN[2]	----/500
POPRES[2]	----/5,919
1000 Genomes Project[3]	----/1,000

1. Illumina HumanHap550.
2. Affymetrix 500K. 3. Whole genome sequence.

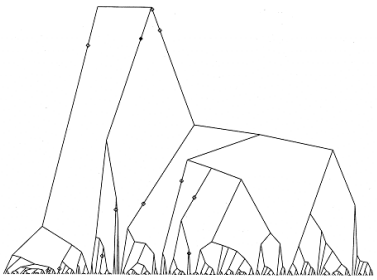
Integration by imputation: Samples genotyped on different genotyping platforms will be integrated by imputing a consistent set of SNPs for all T1D loci.[16-22] To do so we will use HapMap, other available datasets (Table 3) genotyped on the relevant platforms, and the 1000 Genomes Project.[23] We will account for ancestry to the extent possible.[21] We have amassed substantial biological resources for this project, as described in Tables 1-3. The elaboration of our local resources for T1D (Tables 1 & 2) is complete; however the public resources are understated.

Dense map: We will use HapMap and 1000 Genomes Project to infer genotypes for most if not all common SNPs, as well as rare SNPs when possible. *Implementation:* As above.

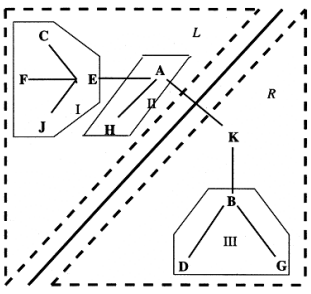
Copy Number Variants (CNVs): CNVs will be called using multiple approaches.[24-26] These findings will help to identify variants that impact genotype calls, are associated as eQTLs, or associate directly with T1D.[27,28]

eQTL: As reviewed in Cookson et al.,[29] eQTL databases, such as VarySysDB[30] and others,[31-33] can be an effective aid to fine-mapping.[33-35] The key idea is to find variants that impact levels of mRNAs plausibly associated with disease.

Single variant (SNP/CNV/etc) association analyses: For the binary outcome of T1D (yes/no) we use logit models with appropriate covariates, such as HLA type, gender, age at onset. We account for subtle variation in ancestry using genetic matching, based on our program Spectral-Gem[36,37] and other approaches.[38,39]



- During the evolutionary history of a T1D region risk variants are embedded in the coalescent tree.
- Haplotypes A-I represent 12 distinct haplotypes (or highly similar haplotypes)
- Haplotypes connected by an edge represent similar haplotypes or clades
  I = (E,C,F,J), II= (A,H), or III = (B, D, G)
- Causal variants are embedded in the history of the haplotype



- Model phenotypes as a function of haplotype effects.
- Cladogram guides the search for causal variants
- E.g., If clade III = (B,D,G) differs strongly from K then the causal variant is embedded in clade III.

Haplotype analyses for association and localization of risk variants: Haplotype-based analyses seek to understand how variation relates to phenotypes and what variation is truly causal. Because causal mutations are embedded in the evolutionary history of the observed haplotypes, these analyses reveal patterns that are more interpretable than direct tests of association between individual variants and phenotype (Figure 2). The approach involves estimating

evolutionary history of the haplotypes to fine map the causal variants.[40-46] Our program eHap implements these evolutionary-based methods and has been modified to utilize haplotypes called by Beagle or Mach.

Model selection: A complementary method to single SNP and haplotypes-based analyses is to find the complete set of SNPs associated with risk variants via model selection. Using model selection procedures for high dimensional linear regression, the lasso[47-58], along with a new statistical approach, "screen and clean", we have developed a method that estimates the true set of genetic variants with nonzero regression coefficients (i.e., the causal SNPs and those in tight LD).[47,57,59] Modifications to accommodate SNPs in tight LD will be incorporated.[58-61] ScreenNclean is available in house and is being constantly enhanced for fine-mapping.

Final product: Integrate results to identify target SNPs for further fine-mapping. It is important to bear in mind that our multiple fine-mapping analyses are targeted at obtaining consistent results and identifying likely risk loci. Issues revolving around multiple testing are not relevant for fine-mapping.

Methodology used during Aim 1b of the project follow:

Statistical models that provide a formal link between SNPs, gene expression, and phenotypes such as T1D are in development.[62-64] These approaches refine signals and help to pinpoint the causal variants.

High Density Genotyping: Genotyping will be performed in the Univ. of Pittsburgh Core facility using the best available 1M platform. We will focus on index cases with early onset, before age 5, and typical HLA risk alleles, starting with cases homozygous for high risk HLA alleles followed by samples from cases heterozygous for one of the high risk haplotypes. We have allocated 10μg of DNA for use during the project. This is sufficient for the genotyping steps, additional materials are available in the event they are needed.

Fit "case-control" models selected in 1a: We will match genetically these genotyped T1D samples to publicly available controls genotyped on a similar platform. Controls will not be screened for T1D, but unscreened controls have little or no impact on power for uncommon diseases (<1%) such as T1D.[65,66] We will then evaluate all models described in Aim 1a, with the goal determining whether the results are consistent with previous findings. We recognize the study is underpowered to discover new loci; our goal is to refine models for target regions to determine how genetic variants affect risk.[67]

Fit models to gene expression and cellular function data: Data on mRNA levels produced under Aims 2-3 will be analyzed to determine its relationship to SNP genotypes (and other variation), while accounting for environmental and experimental covariates. Our approach will use linear models and related Bayesian approaches. Haplotype and model-selection procedures described previously were developed in this linear models setting, and thus are natural approaches for analyses. Certain elements of the genome can affect expression of multiple genes; alternatively LD can cause the illusion of pleiotropy. We will analyze LD structure, haplotypes, multi-SNP models to arrive at a parsimonious solution. Methods will be similar to those in Plagnol et al.[71] We will also guard against false positives due to confounding of genetic variation and gene expression due to differential hybridization differences due to the variation falling in the mRNA region targeted by the microarray probes by using the methods of Alberts et al.[68]

Additional benefit: Perform GWAS on gene expression data to identify novel loci for T1D risk and contribute to the literature on the determinants of gene expression. Identifying QTL affecting gene expression especially relevant to T1D will benefit the field and lay the groundwork for discovering other T1D loci.

Final product: The experiments continue to refine the SNP list for further fine-mapping. At this point we would be ready for targeted genotyping in the samples we have accumulated.

Methodology used during Aim 1c of the project follow:

Genotype selected SNPs from Aim 1b: We expect to have at most 20 SNPs, on average, for each of the T1D loci by this stage of the fine-mapping process. We will implement a staged design to limit expenses while maintaining ability to discriminate models. In Stage 1 we will genotype 2000 cases and 2000 controls on the Illumina 768 SNP BeadArray. In addition to ≈ 600 SNPs for fine-mapping the array will include ≈ 168 SNPs that

are Ancestry Informative Markers or AIMs. The AIMs will be selected from panels identified based on populations of European ancestry and are powerful means of distinguishing different ancestries within Europe.[69,70] If possible additional SNPs will be added to confirm and extend any novel findings from 1a/1b. We expect to have eliminated many SNPs in Stage 1, so that at most an average of 7 SNPs remain as potential risk SNPs at each of the 30 loci. In Stage 2 we will genotype another 2000 cases and 2000 controls on the Illumina 384 SNP BeadArray, again including the same AIMs, to evaluate these SNPs. After this stage it is possible we will have all the data we will need. If not, in Stage 3 we will genotype 40 SNPs in the rest of the samples (6457 cases, many in multiplex families, as well as 2350 controls) using Sequenom technology available at the Univ. of Pittsburgh Core Laboratories.

Fit models selected in 1a/b: With massive samples and appropriate covariates we will evaluate fit of models developed in 1a/b, as well as resolving any inconsistencies by further exploration of the data.

Final product: We expect to determine causal models for a large fraction of T1D loci. It is possible that a few loci will have ambiguities due to loci being in tight LD, as well as expression patterns that are not decipherable.

Aim 2. Characterize the effects of activating cytokines on antigen presenting cells, B-LCLs.

2a) Using B-LCLs analyzed during Aim 1b determine whether changes in cell surface abundance of diabetogenic HLA-DRB1 and -DQB1 alleles exhibit responses to B-cell activating cytokines (IL-4 and IFNγ).

Rationale 2a: The human Major Histocompatibility Complex (MHC) is a highly polymorphic genomic region occupying 4 Mb on chromosome 6p21.[72] Antigen presentation by class I and class II molecules play an important role in controlling immunity and autoimmunity. Precise regulation of MHC molecule expression is critical.[73] Recent data confirm that HLA-B and HLA-A associate with T1D independently of the class II genes HLA-DRB1 and HLA-DQB1.[74] The largest contribution from a single locus (IDDM1) comes from several genes located in the MHC complex, accounting for at least 40% of the familial aggregation.[75] MHC molecules are of central importance for adaptive immunity. The level of MHC molecules expression reveals a quantitatively modulated pattern and directly influences T-cell activation. Tight regulation of MHC class II expression is crucial for the control of the immune response.[76]

Regulation of constitutive and cytokine induced MHC expression is controlled predominantly at the transcriptional level and directly related to both MHC class I and class II gene regulation.[73,77-79] MHC class I and class II molecule expression is stimulated by IFNγ.[77] MHC molecule expression is also modulated by other agents, such as, IL-4, IL-10, IFN-alpha and beta, and TNFα.[73,80]

There is the important question of whether different MHC class isotypes exert distinct T-cell functions ("suppression" of T-cell responses). This issue is of crucial importance in view of the association of MHC class I and class II loci with T1D. Data from HLA fine mapping/GWAS will be combined with precise information on MHC class I and class II molecule constitutive expression and IFNγ stimulation and how MHC expression can be modulated by the Th2-cytokines IL-4/IL-10 analysis will address both (a) qualitative and (b) the quantitatively pattern of expression.

2b) Using an Affymetrix human exon array quantify changes in mRNA abundance, from known and predicted transcribed regions of the entire genome, obtained from resting B-LCL evaluated during Aim 2a.

Rationale 2b: Variation in gene transcription is important in mediating disease susceptibility. Global identification of genetic variants that regulate gene transcription will be helpful in mapping human disease genes.[29] GWAS studies have identified multiple genetic variants that are associated with multifactorial traits.[81] These variants often reside outside of coding regions and will have no known or evident functional effects.[81,82] Gene transcript abundance is directly modified by polymorphism in regulatory elements and consequently may be mapped with considerable power.[29,33] The simultaneous genome-wide assay of gene expression and genetic variation allows the mapping of the genetic factors that underpin individual differences in eQTLs.

Starting point for eQTL mapping is the measurement of gene expression in a target cell system (B-LCLs) from a large number of individuals. This information is the substrate for investigating the effects of DNA polymorphism on the expression of individual genes. For example, a family study of B-LCLs identified nearly 15,000 traits (with each corresponding to an individual Affymetrix probe) with an estimated H^2 >0.3, indicating that genetic influences on gene expression seem to be widespread.[29,83-87]

Methodology used during Aim 2a of the project follow:

Isolation and transformation of PBMCs: PBMCs have already been isolated from whole blood by ficoll-hypaque density gradient centrifugation (Table 2). Briefly, B-LCLs were generated from heparinized plasma-reduced blood diluted with HANK's solution (1:2 dilution). Then 15 mL Ficoll was covered with a layer of diluted blood (30 mL). After 30 min of centrifugation (2000rpm), the PBMC layer was collected. After two washing steps and cell counting, the PBMCs were ready for transformation with EBV. EBV was added directly after isolation of the PBMCs; as an alternative isolated PBMCs were frozen and stored in liquid nitrogen for batch wise transformation. PBMCs were frozen in FBS containing 10% dimethylsulfoxide (DMSO). Freshly isolated or thawed and subsequently washed PBMCs were suspended in 14 mL complete medium (RPMI 1640) with Glutamax, endotoxin free, 10% heat-inactivated fetal bovine serum (FBS, Gibco), 1% penicillin-streptomycin; 0.5% normocin, centrifuged at 350xg for 10 min. The cells were resuspended in 2.5 mL EBV supernatant and 2.5 mL of complete medium, mixed, incubated for at least 3h (37°C; 6% CO_2), and transferred to a 25 cm^2 tissue culture flask. The empty 15 mL Falcon tube was rinsed with 5 mL of cyclosporine (CSA) containing medium before transferring the CSA medium to the cells in the flask; the 10 mL containing flask placed in the incubator. CSA at a final concentration of 1 mg/mL was used to suppress T-cell growth. Flasks were incubated in a humidified incubator at 37°C, 5% CO_2 during a culture period of 28-35 days. Dependent on the cell amount and the quality of the medium the cell cultures were split and expanded until >$2x10^7$ cells were achieved.

Stimulation of B-LCLs: B-LCL cultures will be grown for 48h either untreated or treated with IFNγ or IFNγ plus IL-4. Cells will be frozen and stored in liquid nitrogen or will be immediately used for mRNA extraction.

Total RNA isolation: Total RNA will be isolated using the total RNA isolation protocol (Qiagen RNeasy Mini Kit). This system is based on silica-membrane RNeasy spin columns which bind 100 µg of RNA. Usually 5–$10x10^6$ cells are lysed; yield of tRNA: PBMC per million cells mean 1.5 µg (range:0.8–2.1), B-LCLs per million cells mean 3.2 (range:2.5–4.5). RNase-free DNase I is used to digest DNA. RNA obtained from resting B-LCL cultures will be used during Affymetrix exon array based studies performed in Aim 2b. RNA isolated from stimulated B-LCLs will also be employed but assayed using quantitative PCR assays designed to investigate select mRNAs that were identified during eQTL analysis performed in Aim 1a.

Micro RNA isolation (to be stored as a back-up for further studies): Micro RNA (miRNA) will be isolated according to the manufacturer's protocol using miRNeasy Mini Kit (Quiagen®). PBMC or B-LCLs will be homogenized in 700 µl QIAzol lysis reagent and incubated at room temperature (RT) for 5 min. After addition of 140 µl chloroform and vigorous shaking for 15 s, the homogenate will be separated into phases by centrifugation (15 min, 12000g, 4°C). RNA partitions to the upper, aqueous phase, while DNA partitions to the interphase, and proteins to the lower organic phase and interphase. The upper, aqueous phase will be collected in a new tube, and 1.5 vol 100% ethanol will be added to provide appropriate binding conditions for all RNA molecules from 18 nucleotides upwards. The sample (up to 700 µl) will then be applied to the RNeasy Mini spin column, where the total RNA binds to the membrane and phenol and other contaminants are efficiently washed away (8000rpm for 15 s). High quality RNA will be eluted in RNase-free water.

Quantitative flow cytometry: Cell Activation is a dynamic process therefore we will analyze both the expression of MHC molecules at the cell surface in qualitative terms and in absolute counts (quantum dot technology). To quantify expression levels of markers after stimulation (HLA-class I molecules, -A and -B; and HLA-class II molecules, -DR, -DQ, and -DP), antibodies bound per cell will be determined using QuantiBRITE phycoerythrine (PE) beads (Becton Dickinson). QuantiBRITE-PE is comprised of 4 precalibrated beads to calibrate the FL2 axis in terms of the amount of PE molecules. QuantiBRITE PE beads are reconstituted using 0.5 mL buffer (PBS, 0.1% NaN_3, 0.1% BSA). The beads will be run with a threshold set on forward scatter (FSC) and with instrument settings for fluorescence and compensation the same as for the cellular assay.

Table 4. MHC surface expression will be studied using these monoclonal antibodies.

MHC	Monoclonal	HLA Epitope
Class I[1]	W6/32.HL; IgG2a	-A,-B,-C
Class II[2]	L243: IgG2a	mature -DR
Class II[3]	TÜ 36: IgG2b	Ii assoc. -DR
Class II[3]	TÜ39: IgG2a	-DRB
Class II[3]	TÜ22; IgG2a	-DQ
Class II	B7/21; IgG2	-DP

1. Ref141. 2. Ref142. 3. Ref143.

The assay will be performed as follows: to evaluate variation in expression of MHC molecules B-LCLs will be activated with IFNγ or IFNγ plus IL-4. After activation (3 Mio/50µl) lymphocytes (source PBMC) or (5 Mio/50µl) B-LCLs will be stained with a panel of monoclonal antibodies (Table 4). Data will be acquired on a LSR SORP (BD Bioscience) and analyzed with FACSDiva software.

Isolation of Dendritic Cells (DC): Besides B-LCLs (used as a surrogate for MHC expression on various cells) which will be studied in a large series (N=600 samples) primary human PBMC have also already been isolated from buffy coats (100 ml of whole blood) by density gradient centrifugation. Like the B-LCL samples, the PBMC materials (Table 2) have already been purified and are available for the project. DCs are a major component of the functional antigen-presenting cells, the DC subsets mDC (CD1c+) and pDC (BDCA-2/CD303+), and B-cells (CD19+) will be positively selected using the respective magnetic cell separation kit (Miltenyi Biotec, Bergisch Gladbach, Germany) according to the manufacturer's protocol. The expected amount of enriched mDC is 1-2.5x10^6 mDC per buffy coat (100ml of whole blood obtained from probands who had performed a brief walk for 10-15 min to increase the yield of DCs and lymphocytes). Note that isolation of DCs will be very restricted (i.e., focusing of index cases with early onset, before age 5, and typical HLA risk alleles, starting with cases homozygous for high risk HLA alleles and followed by samples from cases heterozygous for one of the high risk haplotypes) since the protocol will require a large amount of material thus using a significant amount of our existing repository of PBMCs. Likewise to B-LCLs DCs will be stimulated to study constitutive and stimulated MHC class I and class II expression. Expression of MHC molecules on DCs will be correlated with MHC expression on B-LCLs.

Methodology used during Aim 2b of the project follow:

Affymetrix Human Exon Array: Using an Affymetrix human exon array quantify changes in mRNA abundance, from known and predicted transcribed regions of the entire genome, obtained from resting B-LCL evaluated during Aim 2a. mRNA will be isolated as described during Aim 2a. Samples will be shipped to the Univ. of Pittsburgh Core Laboratory for analysis.

Aim 3. Characterize the effect of autoantigen derived peptides on stimulation of memory T-cells.

3) Using PBMCs obtained from cases and controls analyzed during Aim 1b quantify the cytokines (i.e., IL-2, IFNγ, IL-17, TGFβ, and IL-10) secreted by memory T-cells in response to stimulation with mitogenic, diabetes-relevant peptides (i.e., those derived from proinsulin, GAD65, IA2), and control peptides (i.e., those derived from ovalbumin).

Rationale: T-cell–mediated loss of pancreatic β-cells is the crucial event in the development of T1D.[88] β-cell destruction is critically dependent on autoimmune T-cells whose antigen-specific receptors recognize β-cell–derived peptides that bind to risk-associated HLA class II molecules.[89] Autoreactive T-cells represent robust memory cells that recognize multiple GAD65 and preproinsulin derived peptides, as a consequence of epitope spreading.[90] Immunogenicity of autoantigen derived peptides is closely related to kinetic stability of the individual peptide-MHC complexes. The interaction of autoantigen-derived peptides with MHC molecules is critical in modulating T-cell responses.[91] Immunodominant peptides usually reveal high affinity binding characteristics compared to non-immunogenic peptides.[92] Therefore, immunodominance in CD4 T-cell responses is primarily due to an intrinsic property of the peptide-MHC class II complexes.

Data from HLA fine mapping combined with functional T-cell studies will provide information on the dependence of disease-associated CD4 T-cell responses to specific peptide (GAD65/proinsulin)-MHC class II complexes. We (Boehm and Trucco), and others, have recently characterized the phenotypic characteristics of disease-associated T-cells in T1D.[93,94] We used fluorescence-activated cell sorter analysis to study surface marker expression on T-cell lines and PBMC specific for two major T1D autoantigens, GAD65 and proinsulin.[93]

The phenotype of circulating memory T-cells from patients with T1D can be distinguished from those of control subjects by their co-expression of CD25 and CD134.

CD8+ cytotoxic T-cells in insulitic lesions in the pancreas of patients at clinical onset of T1D have been implicated in the disease process.[95] CD8(+) T-cells contribute to a strong IFNγ reactivity against preproinsulin peptides in human T1D. Investigations defining epitope specificity, cytokine secretion, and cytotoxic capacity are important to clarify their role in T1D development. Data show that MHC class I restricted auto-reactive cytotoxic T-lymphocytes (CTLs) are present in the circulation of patients with T1D and that they can kill human islet β-cells.[96] CD8+ T-cells do secrete pro-inflammatory cytokines.[97]

Data from HLA class I fine mapping combined with functional T-cell studies (CD8+) will provide information on the functional properties of auto-reactive CD8+ T-cells and the impact of specific peptide (GAD65/proinsulin)-MHC class I complexes.

The importance of regulatory T lymphocytes (Tregs) in the control of autoimmunity is well established.[98] The emergence of Tregs as an essential component of immune homeostasis provides the opportunity to study the number or the lack of T-cells with a regulatory phenotype in the context of T1D risk genes. Resistant Treg memory T-cells and/or "defective" Treg favor a breakdown in tolerance. Tregs have recently been purified: CD4(+)CD127(lo/-) and CD4(+)CD127(lo/-)CD25(+) T-cells. CD4(+)CD127(lo/-)CD25(+) T-cells and CD45RA(+) subset are functional Tregs with higher level of FOXP3 expression.[99] Data from GWAS/expression profiling will be combined with phenotypic characteristics from multi-parameter FACS analysis. FACS analysis will focus on (a) Tregs and (b) expression of co-stimulatory signals[100] and effector function of CD8 T-cells.

Methodology used during Aim 3 of the project follow:

T-cell stimulation: Assays will be set up in 96-well flat-bottom microtiter plates using $4x10^5$ PBMC/well in a volume of 200 µl. The PBMC will be incubated in culture medium (RPMI including Glutamax, 10% human serum, 1% penicillin/streptomycin) with or without antigen at 37°C/7% CO_2. Incubation time is dependent on the assay and will range from 24h (intracellular cytokine staining) up to 120h (expression of surface markers). Antigens are peptide-panels from the major autoantigens Insulin/Proinsulin/GAD65 and IA-2 used at a final concentration of 10 µg/ml. Positive controls for intracellular staining assays are the mitogen phytohemagglutinin (1 µg/ml) or staphylococcal enterotoxin B (1 µg/ml). Control antigens for all immunological assays are Tetanus Toxoid (5 µg/ml), Ovalbumin (5 µg/ml) and purified protein derivative (5 µg/ml).

Cytokine assays: The supernatants will be collected and stored in aliquots at -80 °C until analysis from all experiments. Cytokines (IL-2, IFNγ, IL-17, TGFβ, and IL-10) will be quantified by ELISA (all from R&D Systems) and/or the cytometric bead array (BD Biosciences) method.

ELISA: All assays are commercially available and will be performed according to manufacturer's instructions. 96-well plates will be incubated overnight at RT with the cytokine-specific capture antibody. After 3 washing steps block buffer will be added to the wells and the plate will be incubated at RT for 1h. After 3 washing steps samples and standards will be added. After an incubation time of 2h at RT plates will be washed, before the cytokine specific detection antibody is added. After 2h of incubation and 3 washing steps, Streptavidin-HRP will be added. The plates will be incubated for 20 min at RT in the dark. After a final washing procedure, substrate solution will be added. After 20 min at RT in the dark, stop solution will be added and subsequently, the optical density will be determined at a certain, cytokine-specific, wavelength. The optical density will be measured and analyzed using EL808 ELISA Reader and the Gen5TM software (BioTek).

Cytokine bead array: The Cytometric Bead Array kit (BD Bioscience) will be used to quantify levels of secreted cytokines in supernatants. Assays will be performed according to manufacturer's instructions and cytokines analyzed using BD Cytometric BeadArray software and the LSR SORP (BD Bioscience).

Multiparameter FACS analysis: Detection of antigen-specific T-cells is difficult due to the little amount of these cells in the periphery (<0.1% lymphocytes). To determine both phenotypic and functional characteristics of these cells, multicolour FACS will be applied. The technique enables distinct measurements performed

simultaneously on a single sample (a must, since we are handling biological/unique samples). Subsets of CD4 as well as CD8 T-cells will be analyzed for their antigen-specific expression profile of activation markers as well as secreted cytokines after stimulation. The effector function or the cytotoxic capacity of CD8 T-cells will also be analyzed. Data will be acquired on a LSR SORP (BD Bioscience) analyzed with FACSDiva software using well-defined gating strategies.

Surface marker phenotyping: Following antigen-specific stimulation, PBMC will be harvested and transferred into V-bottom 96-well plates. The transferred cells will be incubated for 30 min at 4°C with the appropriate staining protocol (Table 5). After a washing step the cells will be directly analyzed by flow cytometry. Both, frequency of cells with specific phenotypes and absolute numbers will be defined.

Table 5. Fluorescent staining protocol.

Conjugate	CD4 Teff	Treg	K^+-channel block	Treg/Teff	CD8
FITC	CD134	CD45RA	Kv3.1	TNFα	CD107a/b
PE	TNFα	FoxP3	CD134	FoxP3/GPR83	IL-2
PE-Cy5	-	-	-	CD134	
PerCP	CD3	CD3	CD3	CD3	CD8
PerCP-Cy5	-	-	-	IFNγ	-
PE-Cy7	CD25	CD127	CD25	CD127/GPR83	IFNγ
APC	IFNγ	IL-10	CCR7	IL-10	CD137
AlexaFlour700	-	CD25	-	CD25	-
APC-Cy7	CD4	CD4	CD4	CD4	CD3
Pacific Blue	CD45RO	-	CD45RO	CD45RO	CD45RO

Intracellular staining: Prior to the intracellular staining procedure, the PBMC will be stimulated with the appropriate antigens overnight (16h, 37°C, 5% CO_2). To keep cytokines in the cell, Brefeldin A (Sigma) will be added after 2h (1 µg/mL). After stimulation overnight, the cells will be harvested and transferred to a V-bottom 96 plate ($1x10^6$cells/well). The supernatants will be decanted following a centrifugation step (1300rpm, 5 min). The cell pellets will then be resuspended with the appropriate volume (50 µl per $1x10^6$ cells) of fixation buffer followed by incubation at RT for 30 min. After centrifugation (1300rpm, 5 min) the supernatants will be discarded and subsequently, the cells will be resuspended with the appropriate volume (50 µl per $1x10^6$ cells) of permeabilization buffer. After centrifugation and decanting of the supernatant, the cells will be incubated for 30 min at RT in the dark with the appropriate staining protocol (Table 5). The cells will be washed twice and subsequently analyzed by flow cytometry.

Aim 4. Identify causal gene networks implicated by results in Aims 1-3 and use those results to search for new T1D risk loci.

Methodology used during Aim 4 of the project follow:

Uncover T1D gene co-expression networks: Results from Aims 1-3 set the stage for identifying networks of genes critical for normal development and T1D risk. Microarray data from Aims 2-3 embody relationships among gene transcripts. Two approaches for finding networks from these data can be dubbed empirical and Bayesian. Both seek to uncover pathways via the correlation in amounts and kinds of mRNA found by the experiment. The empirical approach discovers gene co-expression networks solely from the structure within the correlation matrix of mRNA levels, such as by eigenvector or cluster analysis.[101,102] Key components of the network can be identified by how the genes, represented as nodes, communicate with other nodes in the network (i.e., measured by strength of correlation of expression levels and related measures); communication between networks or modules[103-106] can be likewise characterized. Our goal is to determine if and how genetic variation falling in our T1D loci plays a role in the networks and communication across networks, as well as searching for novel genetic variation from T1D GWAS studies.[107]

Early models treated genes equally, a priori, when testing genes or sets of genes for association, but this approach ignores the important information available concerning gene networks that reflect the fact that neighboring genes in the network have shared biological functions. We plan to follow the Bayesian approach developed by W. Pan which utilizes a prior on network structure to enhance power.[62-64]

Discover new T1D risk variants: Networks give us a new tool to re-explore all GWAS data, including data produced by Aim 1, to discover new risk variation for T1D. We will use the following approaches.

Weighted analyses: One approach to analysis of networks builds on ideas that Devlin and Roeder introduced into the literature for genetic association tests.[108,109] Differentially weighting groups of genes, with weights determined by the data, improves the power to detect signals in likely groupings, while maintaining controls of false positives. Remarkably this approach is effective even if only a small fraction of the groupings are informative. Others have built on these ideas,[110-112] including models that include prior knowledge about genetic pathways or networks to form a Bayesian model[62-64] or by building graphical models.[113,114] Both of these approaches seem very promising and we plan to pursue them.

Extended haplotypes-sharing to detect rare variants: Natural selection alters more than the allele frequency of the selected genomic variant. It alters frequencies for adjacent loci, and often leaves a trail of "extended haplotypes" around the selected variant. In fact such extended haplotypes have even been used to find eQTLs.[115] We plan to use and refine methods for finding extended haplotypes. We will examine the data in multiple ways to discover regions exhibiting extended haplotypes, in this case associated with rare risk alleles. We view the haplotype analysis as hypothesis generating, requiring further study for confirmation. One approach we (Devlin and Roeder) helped to pioneer reduces dimensionality of the tests by evaluating "haplotype matching" among pairs of individuals[116,117] The test is powerful for detecting multiple rare mutations. A complementary method in PLINK finds unusual stretches of homozygosity in cases versus controls.[25] PLINK can also detect extended IBD-sharing between pairs of distantly related individuals by use of a hidden Markov model. The procedure also provides a test for association.

Confirm variants in 10,457 cases and 6750 controls: Again we will implement a staged design to limit expenses while maintaining ability to discriminate models. In Stage 1 we will genotype 2000 cases and 2000 controls on the Illumina 384 SNP BeadArray, and will include some European AIMs. Then, having eliminated many loci as potential risk loci, in Stage 2 we will genotype another 2000 cases and 2000 controls for 40 SNPs for the rest of the samples using Sequenom technology. See 1c for statistical approach.

Relate genetic and expression data to determine function of new risk loci: Please see Aim 1b/c, 2, and 3.

Goal 1. INNOVATION

Results from the project will fully define the LD structure of each T1D-associated genomic region; identify T1D risk variants; determine their method of action; and estimate the effect on risk for the polymorphisms. eQTLs have been described for T1D (e.g., *HLA-DQB*0302* expression in B-LCLs, *insulin* gene expression in the thymus and differential splicing of *CTLA4*)[119]. Network analysis has been applied to understand the context in which genes operate and how context influences T1D risk, using the data obtained from primary human hepatocytes to connect the Chromosome 12q13 risk variant with *RPS26* expression.[118] Although these data have been challenged,[71] we believe there are true eQTL signals, as observed for *HLA-DQB*, *INS* and *CTLA4*, associated with T1D. The goal is the identification of genes in the context of their functional network that define etiologic pathways. An advantage to our study design is that data will be obtained from antigen presenting cell surrogates and autoreactive T-cells, two tissues implicated in T1D pathogenesis. Knowledge of the context in which genes operate is essential for development of new therapeutic targets.

There are a number of innovations that will result from the research. During the course of our work we will isolate total mRNA for use in the current project but will also isolate miRNAs in order to create a new repository for use in evaluating the role of small interfering RNAs. Efforts resulting from the work of our statistical group will result in novel methodologies and new theoretical approaches to gene discovery. We anticipate the project will provide results of intensive genetic, gene expression, and cellular functional studies from ancestry matched cases and controls. The data will be available via appropriate public resources (e.g., dbGAP).

Goal 1. INVESTIGATOR(S) QUALIFICATIONS.

Massimo Trucco, MD, PI, will be responsible for administration of the project. His laboratory will work with the research teams headed by our collaborators. Dr. Trucco has recruited Dr. Ringquist, an Assistant Professor at Children's Hospital of Pittsburgh, to coordinate interactions between the various research groups (Ringquist already serves in this capacity in an earlier, collaboration with Devlin and Roeder). The tasks that will be assigned to the Trucco laboratory are protocols in Aims 2-3. This will be done in collaboration with Boehm's

group with the Trucco laboratory working with U.S. samples and Boehm's group focusing on samples from Europe. To recognize and avoid systematic errors that can occur in different laboratories the two groups will stay in frequent contact sharing data and materials. We have permission from our IRB to transfer samples from Germany to Pittsburgh and to exchange electronic data in both directions, and including Devlin and Roeder as well. Trucco's laboratory has expertise applying molecular techniques to study genetic problems, specifically clarifying the molecular basis of *HLA-DQ* polymorphism in T1D using B-LCLs generated from families of diabetic probands.[120,121] Trucco has published extensively with the team members involved in this project, resulting in 13 publications.[36,122-133] In addition to directing his own research laboratory Trucco has served as the Director the Histocompatibility Laboratory for Tissue Typing (Univ. of Pittsburgh) and as the Director Children's Hospital of Pittsburgh Histocompatibility Center. The former laboratory is dedicated to selection of donor/recipient pairs for kidney, pancreas, heart, liver and lung transplantation. The latter laboratory annually molecularly typed and stored ~200,000 blood samples from bone marrow donors from the entire U.S. Trucco has received both the William Stadie Award of the American Diabetes Association and the Univ. of Michigan Sandoz Prize, for the study–conducted in collaboration with Dr. Starzl–of immunologic microchimerism in transplanted patients.[134,135] Dr. Trucco has also received the Univ. of Pittsburgh Chancellor's Distinguished Award for his work on the etiology of T1D,[136,137] constituting the basis for the Children's Hospital of Pittsburgh Diabetes Institute (Est., 2002), of which he became the first Scientific Director.

Dr. Boehm has worked for more than 20 years in the field of T1D. Since 1988 Dr. Boehm has been the PI in population based studies dealing with both overall prevalence and genetic basis of islet cell autoimmunity, the Ulm School-Children Study.[138] Boehm`s laboratory is acting as the central lab facility for the European branch of the multinational NIH-funded T1D Genetics Consortium (T1DGC), which has recruited more than 1,500 multiplex T1D families in Europe. Dr. Boehm`s lab has expertise in establishing growth and functional characterization of B-LCLs, and in phenotypic, functional and molecular characterization of autoreactive T-cells.[89,93,133] The impact of genetic polymorphisms to the T-cell cytokine signature has been extensively studied by Boehm`s group.[97,139,140] Dr. Boehm is professor of medicine at Ulm Univ., Germany since 1993, where he serves as the Medical Director of the Division of Endocrinology, Diabetes and Metabolism. Since 2005 he has also taken over the position of the Director of the Centre of Excellence for Metabolic Diseases at Ulm Univ. and he is acting as the Vice-President of Ulm International Graduate School. Since 1989 the Trucco and Boehm laboratories have had a very active collaboration, resulting in 12 publications[122-133].

Devlin and Roeder have extensive experience in modeling genetic data, dating back over 20 years. They have contributed to the theoretical framework underlying our present understanding of LD and haplotype structure and how the evolution of haplotypes can be used in fine mapping. In addition they have developed novel statistical methodologies for combining multiple types of data, and biological pathway information, to increase statistical power and refinement of signals. Recently they have developed model selection methods that facilitate the search for associated variants and interactions that are much more sensitive to detecting causal variants than traditional tests that consider SNPs individually. Devlin coordinates much of the data analysis at the Univ. of Pittsburgh School of Medicine. Devlin's group has experience handling massive datasets, including leadership of the analysis team for the Autism Genome Project. The Department of Human Genetics is next door, and Devlin is an adjunct member. Roeder is a Professor of Statistics at Carnegie Mellon Univ. (CMU). CMU has one of the finest statistics departments in the country. It specializes in computationally intensive methods for data analysis and features the Department of Machine Learning, a consortium of faculty members from the Departments of Statistics and Computer Science. Several of Roeder's colleagues also work in statistical genetics, including Dr. Howard Seltman and Dr. Larry Wasserman both of whom are supported under MH057881, which supports the bulk of the theory research for the Devlin and Roeder groups. We find that new methodological work often springs from the demands of genetic data, and we expect this project to inspire new ways of fine-mapping risk loci for T1D and human diseases more generally.

Although we have worked together for a number of years with little formal coordination we have, due to the intensive nature of this project, chosen to adopt a formalized means of coordination. E-mail is a quick and effective means of communication, but we also plan conference calls bimonthly to ensure research momentum and resolve challenges, as well as yearly meetings to solidify the collaboration. The decision-making body is the entire group of investigators, who will each have one vote. We will not need formal subcommittee structure. Expertise has determined where the principal efforts are exerted. Authorship will be determined by

contribution. We anticipate debate about scientific approaches, but do not anticipate any significant differences about science or process. All issues will be resolved by majority-rule vote. Data management will be the initial responsibility of the investigator. However Devlin's group will integrate the data sets and maintain an integrated database. The database will reside on a Univ. of Pittsburgh mass storage device, and backed-up nightly. Devlin will be responsible for sharing the data with other investigators. Except in special cases, transfer of de-identified data will be via the web, between password-protected hidden sites.

Goal 1. TIMELINE

Yr. 1: Fine-mapping by meta-analysis of publicly-available GWAS; Initiate GWAS genotyping, B-LCL growth assays and characterization of memory T-cell populations; Initiate analysis of gene expression

Yr. 2: Further fine-mapping and GWAS studies; Continue B-LCL growth assays, gene expression analysis, and characterization of memory T-cell populations

Yr. 3: Completion of GWAS studies and Stage 2 fine-mapping; Completion of B-LCL assays, analysis of gene expression, and T-cell studies; Discovery of eQTLs

Yr 4: Integrative analysis of eQTLs followed by cellular functional data and gene pathway informatics to identify causal T1D gene networks; Initiate study of selected eQTLs using materials obtained from affected sib-pairs.

Yr 5: Completion of studies with affected sib-pairs; Refinement of gene network models; Confirm select interactions via cellular studies with B-LCL and memory T-cells obtained from selected study participants.

Goal 1. LITERATURE CITATIONS:

1. Smyth DJ, Plagnol V, Walker NM, Cooper JD, Downes K, Yang JH, Howson JM, Stevens H, McManus R, Wijmenga C, Heap GA, Dubois PC, Clayton DG, Hunt KA, van Heel DA, Todd JA. Shared and distinct genetic variants in type 1 diabetes and celiac disease. N Engl J Med 359:2767-2777, 2008.

2. Silveira PA, Grey ST. B cells in the spotlight: innocent bystanders or major players in the pathogenesis of type 1 diabetes. Trends Endocrinol Metab 17:128-135, 2006.

3. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678, 2007.

4. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, Lowe CE, Szeszko JS, Hafler JP, Zeitels L, Yang JH, Vella A, Nutland S, Stevens HE, Schuilenburg H, Coleman G, Maisuria M, Meadows W, Smink LJ, Healy B, Burren OS, Lam AA, Ovington NR, Allen J, Adlem E, Leung HT, Wallace C, Howson JM, Guja C, Ionescu-Tîrgovişte C; Genetics of Type 1 Diabetes in Finland, Simmonds MJ, Heward JM, Gough SC; Wellcome Trust Case Control Consortium, Dunger DB, Wicker LS, Clayton DG. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39:857-864, 2007.

5. Hakonarson H, Grant SF, Bradfield JP, Marchand L, Kim CE, Glessner JT, Grabs R, Casalunovo T, Taback SP, Frackelton EC, Lawson ML, Robinson LJ, Skraban R, Lu Y, Chiavacci RM, Stanley CA, Kirsch SE, Rappaport EF, Orange JS, Monos DS, Devoto M, Qu HQ, Polychronakos C. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature 448:591-594, 2007.

6. Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, Allen JE, Downes K, Barrett JC, Healy BC, Mychaleckyj JC, Warram JH, Todd JA. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. Nat Genet 40:1399-1401, 2008.

7. Hakonarson H, Qu HQ, Bradfield JP, Marchand L, Kim CE, Glessner JT, Grabs R, Casalunovo T, Taback SP, Frackelton EC, Eckert AW, Annaiah K, Lawson ML, Otieno FG, Santa E, Shaner JL, Smith RM, Onyiah CC, Skraban R, Chiavacci RM, Robinson LJ, Stanley CA, Kirsch SE, Devoto M, Monos DS, Grant SF,

Polychronakos C. A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. Diabetes 57:1143-1146, 2008.

8. Grant SF, Qu HQ, Bradfield JP, Marchand L, Kim CE, Glessner JT, Grabs R, Taback SP, Frackelton EC, Eckert AW, Annaiah K, Lawson ML, Otieno FG, Santa E, Shaner JL, Smith RM, Skraban R, Imielinski M, Chiavacci RM, Grundmeier RW, Stanley CA, Kirsch SE, Waggott D, Paterson AD, Monos DS; DCCT/EDIC Research Group, Polychronakos C, Hakonarson H. Follow-up analysis of genome-wide association data identifies novel loci for type 1 diabetes. Diabetes 58:290-295, 2009.

9. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 322:881-888, 2008.

10. Holl RW, Boehm B, Loos U, Grabert M, Heinze E, Homoki J. Thyroid autoimmunity in children and adolescents with type 1 diabetes mellitus. Effect of age, gender and HLA type. Horm Res 52:113-118, 1999.

11. Hunt KA, Zhernakova A, Turner G, Heap GA, Franke L, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, Panesar D, Gwilliam R, Takeuchi F, McLaren WM, Holmes GK, Howdle PD, Walters JR, Sanders DS, Playford RJ, Trynka G, Mulder CJ, Mearin ML, Verbeek WH, Trimble V, Stevens FM, O'Morain C, Kennedy NP, Kelleher D, Pennington DJ, Strachan DP, McArdle WL, Mein CA, Wapenaar MC, Deloukas P, McGinnis R, McManus R, Wijmenga C, van Heel DA. Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet 40:395-402, 2008.

12. Frazer KA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861, 2007.

13. Sabeti PC, et al. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913-918, 2007.

14. Rinaldo A, Bacanu SA, Devlin B, Sonpar V, Wasserman L, Roeder K. Characterization of multilocus linkage disequilibrium. Genet Epidemiol 28:193-206, 2005.

15. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265, 2005.

16. Zaitlen N, Kang HM, Eskin E, Halperin E. Leveraging the HapMap correlation structure in association studies. Am J Hum Genet 80:683-691, 2007.

17. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906-913, 2007.

18. Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3:e114, 2007.

19. Nicolae DL. Testing untyped alleles (TUNA)-applications to genome-wide association studies. Genet Epidemiol 30:718-727, 2006.

20. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210-223, 2009.

21. Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P. Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet 84:235-250, 2009.

22. www.sph.umich.edu/csg/abecasis/mach/

23. www.1000genomes.org/

24. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17:1665-1674, 2007.

25. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40:1253-1260, 2008.

26. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35:2013-2025, 2007.

27. Stranger B, Forrest M, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853, 2007.

28. Stranger B, Nica A, Forrest M, et al. Population genomics of human gene expression. Nat Genet 39:1217–1224, 2007.

29. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet 10:184-194, 2009.

30. Yamasaki C, Murakami K, Fujii Y, Sato Y, Harada E, Takeda J, Taniya T, Sakate R, Kikugawa S, Shimada M, et al. The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. Nucleic Acids Res 36:D793–D799, 2008.

31. Choy E, Yelensky R, Bonakdar S, Plenge RM, Saxena R, De Jager PL, Shaw SY, Wolfish CS, Slavik JM, Cotsapas C, Rivas M, Dermitzakis ET, Cahir-McFarland E, Kieff E, Hafler D, Daly MJ, Altshuler D. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. PLoS Genet 4:e1000287, 2008.

32. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4:e1000214, 2008.

33. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO. A genome-wide association study of global gene expression. Nat Genet 39:1202-1207, 2007.

34. Graham, RR; Kozyrev, SV; Baechler, EC; Reddy, MV; Plenge, RM, et al. A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. Nat Genet 38:550–555, 2006.

35. Moffatt, MF; Kabesch, M; Liang, L; Dixon, AL; Strachan, D, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448:470–473, 2007.

36. Luca D, Ringquist R, Klei L, Lee AB, Gieger C, Wichmann H-E, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82:453-463, 2008.

37. Lee WC. Case-control association studies with matching and genomic controlling. Genet Epidemiol 27:1-13, 2004.

38. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904-909, 2006.

39. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190, 2006.

40. Templeton AR. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E locus. Genetics 140:403–409, 1995.

41. Templeton AR, Boerwinkle E, Sing CF. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. Genetics 117:343–351, 1987.

42. Templeton AR, Crandall KA, Sing CF. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics 132:619–633, 1992.

43. Templeton AR, Sing CF, Kessling A, Humphries S. A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. Genetics 120:1145–1154, 1988.

44. Templeton AR, Sing CF. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. Genetics 134:659–669, 1993.

45 Seltman H, Roeder K, Devlin B. TDT meets MHA: Family-based association analysis guided by the evolution of haplotypes. Am J Hum Genet 68:1250-1263, 2001.

46. Seltman, H, Roeder K, Devlin B. Evolutionary-based association analysis using haplotype data. Genet Epidemiol, 25:48-58, 2003.

47. Wasserman L, Roeder K. High Dimensional Variable Selection. Ann Stat, in press (& online), 2008.

48. Meinshausen N, Buhlmann, P. High dimensional graphs and variable selection with the lasso. The Annals of Statistics 34:1436–1462, 2006

49. Candes, E, Tao, T. The Dantzig selector: statistical estimation when p is much larger than n %arxiv.org/math.ST/0506081, 2005.

50. Wainright, M. Sharp thresholds for high-dimensional and noisy recovery of sparsity. arxiv.org/math.ST/0605740, 2006

51. Zhao P, Yu B. On model selection consistency of lasso. Journal of Machine Learning Research 7:2541-2563, 2006.

52. Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101:1418-1429, 2005.

53. Fan, J, Lv, J. Sure independence screening for ultra-high dimensional feature space. Manuscript, 2006.

54. Meinshausen N, Yu B. Lasso-type recovery of sparse representations of high-dimensional data. Technical report. Berkeley, 2006.

55. Tropp JA. Greed is good: algorithmic results for sparse approximation. IEEE Transactions on Information Theory 50:2231-2242, 2004.

56. Tropp JA. Just relax: convex programming methods for identifying sparse signals in noise IEEE Transactions on Information Theory 52:1030-1051, 2006.

57. Meinshausen, N. Lasso with relaxation. Computational Statistics and Data Analysis, 2006.

58. Meinshausen N, Meier L, Buhlmann P. P-values for high-dimensional regression. Tech rep. Seminar fur Statistik. ETH Zurich, 2008.

59. Wu J, Devlin B, Roeder K. Screen and Clean: a tool for identifying interactions in genome-wide association studies. Unpublished.

60. Sperrin M, Jaki T. Direct Effects Testing: A two-stage procedure to test for effect size and variable importance for correlated binary predictors and a binary response. Unpublished manuscript.

61. Malo N, Libiger O, Schork NJ. Accommodating linkage disequilibrium in genetic association analyses via ridge regression. American Journal of Human Genetics, in press.

62. Pan W, Jeong KS, Xie Y, Khodursky A. A nonparametric empirical Bayes approach to joint modeling of multiple sources of genomic data. Statistica Sinica 18:709-729, 2008.

63. Wei P, Pan W. Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. Bioinformatics 24:404-411, 2008.

64. Pan W. Incorporating gene functional annotations in detecting differential gene expression. Appl Statist. 55:301-316, 2006.

65. Bacanu S-A., Devlin B, Roeder K. The power of genomic control. Am J Hum Genet 66:933-944, 2000.

66. Li M, Boehnke M, Abecasis GR. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. Am J Hum Genet 78:778-792, 2006.

67. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer, NY, 2001.

68. Alberts R, Terpstra P, Li Y, Breitling R, Nap JP, Jansen RC. Sequence polymorphisms cause many false cis eQTLs. PLoS ONE 2:e622, 2007.

69. Price AL, Butler J, Patterson N, Capelli C, Pascali VL, et al. Discerning the ancestry of European Americans in genetic association studies. PLoS Genet doi: 10.1371/journal.pgen.0030236, 2008.

70. Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, et al. Analysis and application of European genetic substructure using 300K SNP information. PLoS Genet doi: 10.1371/journal.pgen.0040004, 2008.

71. Plagnol V, Smyth DJ, Todd JA, Clayton DG. Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. Biostatistics 10:327-334, 2009.

72. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC Jr, Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S. Gene map of the extended human MHC. Nat Rev Genet 5:889-899, 2004.

73. Ting JP, Trowsdale J. Genetic control of MHC class II expression. Cell 109:S21-33, 2002.

74. Badenhoop K, Kahles H, Seidl C, Kordonouri O, Lopez ER, Walter M, Rosinger S, Ziegler A, Boehm BO; Diabetes Genetics Consortium. MHC-environment interactions leading to type 1 diabetes: feasibility of an analysis of HLA DR-DQ alleles in relation to manifestation periods and dates of birth. Diabetes Obes Metab 11:31-45, 2009.

75. Thomson G, Valdes AM, Noble JA, Kockum I, Grote MN, Najman J, Erlich HA, Cucca F, Pugliese A, Steenkiste A, Dorman JS, Caillat-Zucman S, Hermann R, Ilonen J, Lambert AP, Bingley PJ, Gillespie KM, Lernmark A, Sanjeevi CB, Rønningen KS, Undlien DE, Thorsby E, Petrone A, Buzzetti R, Koeleman BP, Roep BO, Saruhan-Direskeneli G, Uyar FA, Günoz H, Gorodezky C, Alaez C, Boehm BO, Mlynarski W, Ikegami H, Berrino M, Fasano ME, Dametto E, Israel S, Brautbar C, Santiago-Cortes A, Frazer de Llado T, She JX, Bugawan TL, Rotter JI, Raffel L, Zeidler A, Leyva-Cobian F, Hawkins BR, Chan SH, Castano L, Pociot F, Nerup J. Relative predispositional effects of HLA class II DRB1-DQB1 haplotypes and genotypes on type 1 diabetes: a meta-analysis. Tissue Antigens 70:110-27, 2007.

76. Drozina G, Kohoutek, Jabrane-Ferrat, Peterlin BM. Expression of MHC II genes. Curr. Top. Microbiol. Immunol 290:147-170, 2005.

77. Beaty JS, West KA, Nepom GT. Functional effects of a natural polymorphism in the transcriptional regulatory sequence of HLA-DQB1. Mol Cell Biol 15:4771-4782, 1995.

78. Beaty JS, Sukiennicki TL, Nepom GT. Allelic variation in transcription modulates MHC class II expression and function. Microbes Infect 1:919-927, 1999.

79. Andersen LC, Beaty JS, Nettles JW, Seyfried CE, Nepom GT, Nepoom BS. Allelic polymorphism in transcriptional regulatory regions of HLA-DQB genes. J Exp Med 173:181-192, 1991.

80. Reith W, LeibundGut-Landmann S, Waldburger JM.Regulation of MHC class II gene expression by the class II transactivator. Nat Rev Immunol 5:793-806, 2005.

81. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118:1590-1605, 2008.

82. Kleinjan DA, van Heyningen V. Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 76:8-32, 2005.

83. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365-1369, 2005.

84. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. Nature 2004 430:743-747, 2004.

85. Hoffmann MM, Winkler K, Renner W, Winkelmann BR, Seelhorst U, Wellnitz B, Boehm BO, März W.Genetic variants and haplotypes of lipoprotein associated phospholipase A2 and their influence on cardiovascular disease (The Ludwigshafen Risk and Cardiovascular Health Study). J Thromb Haemost 7:41-48, 2009.

86. Grammer TB, März W, Renner W, Boehm BO, Hoffmann MM. C-reactive protein genotypes associated with circulating C-reactive protein but not with angiographic coronary artery disease: the LURIC study. Eur Heart J. 30:170-182, 2009.

87. Winkelmann BR, Hoffmann MM, Nauck M, Kumar AM, Nandabalan K, Judson RS, Boehm BO, Tall AR, Ruaño G, März W. Haplotypes of the cholesteryl ester transfer protein gene predict lipid-modifying response to statin therapy. Pharmacogenomics J 3:284-296, 2003.

88. Trucco M. Regeneration of the b cell. J Clin Invest 115:5, 2005.

89. Durinovic-Belló I, Rosinger S, Olson JA, Congia M, Ahmad RC, Rickert M, Hampl J, Kalbacher H, Drijfhout JW, Mellins ED, Al Dahouk S, Kamradt T, Maeurer MJ, Nhan C, Roep BO, Boehm BO, Polychronakos C, Nepom GT, Karges W, McDevitt HO, Sønderstrup G. DRB1*0401-restricted human T cell clone specific for the

major proinsulin73-90 epitope expresses a down-regulatory T helper 2 phenotype. PNAS 103:11683-11688, 2006.

90. Lehmann PV, Forsthuber T, Miller A, Sercarz EE. Spreading of T-cell autoimmunity to cryptic determinants of an autoantigen. Nature 358:155–157, 1992.

91. Trucco M, Giannoukakis N. MHC tailored for diabetes cell therapy. Gene Therapy 12:553, 2005.

92. Sant AJ, Chaves FA, Jenks SA, Richards KA, Menges P, Weaver JM, Lazarski CA. The relationship between immunodominance, DM editing, and the kinetic stability of MHC class II:peptide complexes. Immunol Rev 207:261-278, 2005.

93. Endl J, Rosinger S, Schwarz B, Friedrich SO, Rothe G, Karges W, Schlosser M, Eiermann T, Schendel DJ, Boehm BO. Differential Roles of Costimulatory Signaling Pathways in Type 1 Diabetes Mellitus. Diabetes 55:50-60, 2006.

94. Luppi P, Zanone MM, Hyoty H, Rudert WA, Haluszczak C, Alexander AM, Bertera S, Becker D, Trucco M. Restricted TCR V beta gene expression and enterovirus infection in type I diabetes: a pilot study. Diabetologia 43:1484-1497, 2000.

95. In't Veld P, Lievens D, De GJ, Ling Z, Van der Auwera B, Pipeleers-Marichal M, Gorus F, Pipeleers D. Screening for insulitis in adult autoantibody-positive organ donors. Diabetes 56:2400–2404, 2007

96. Skowera A, Ellis RJ, Varela-Calviño R, Arif S, Huang GC, Van-Krinks C, Zaremba A, Rackham C, Allen JS, Tree TI, Zhao M, Dayan CM, Sewell AK, Unger W, Drijfhout JW, Ossendorp F, Roep BO, Peakman M. CTLs are targeted to kill beta cells in patients with type 1 diabetes through recognition of a glucose-regulated preproinsulin epitope. J Clin Invest 118:3390-3402, 2008.

97. Durinovic-Belló I, Schlosser M, Riedl M, Maisel N, Rosinger S, Kalbacher H, Deeg M, Ziegler M, Elliott J, Roep BO, Karges W, Boehm BO. Pro- and anti-inflammatory cytokine production by autoimmune T cells against preproinsulin in HLA-DRB1*04, DQ8 Type 1 diabetes. Diabetologia 47:439-450, 2004.

98. Brusko TM, Putnam AL, Bluestone JA. Human regulatory T cells: role in autoimmune disease and therapeutic opportunities. Immunol Rev 223:371-390, 2008.

99. Putnam AL, Brusko TM, Lee MR, Liu W, Szot GL, Ghosh T, Atkinson MA, Bluestone JA. Expansion of human regulatory T-cells from patients with type 1 diabetes. Diabetes 58:652-662, 2009.

100. Boehm BO, Bluestone JA. Differential roles of costimulatory signaling pathways in type 1 diabetes mellitus. Rev Diabet Stud. 1:156-164, 2004.

101. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 1:54, 2007.

102. Aten JE, Fuller TF, Lusis AJ, Horvath S. Using genetic markers to orient the edges in quantitative trait networks: the NEO software. BMC Systems Biology 2:34, 2008.

103. Albert R. Scale-free networks in cell biology. J Cell Sci 118:4947-4957, 2005.

104. Barabási A, Oltvai Z. Network Biology: Understanding the Cell's Functional Organization. Nature Reviews: Genetics 5:101-113, 2004.

105. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34:166-176, 2003.

106. Carter S, Brechb C, Griffin M, Bond A. Gene Co-expression Network Topology Provides a Framework for Molecular Characterization of Cellular State. Bioinformatics 20:2242-2250, 2004.

107. www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/

108. Roeder K, Bacanu SA, Wasserman L, Devlin B. Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet 78:243-252, 2006.

109. Roeder K, Devlin B, Wasserman L. Improving power in genome-wide association studies: weights tip the scale. Genet Epidemiol 31:741-747, 2007.

110. Ionita-Laza I, McQueen MB, Laird NM, Lange C. Genomewide Weighted Hypothesis Testing in Family-Based Association Studies, with an Application to a 100K Scan. Am J Hum Genet 81:607-614, 2007.

111. Sun L, Craiu RV, Paterson AD, Bull SB. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. Genet Epidemiol 30:519-530, 2006.

112. Greenwood CM, Rangrej J, Sun L. Optimal selection of markers for validation or replication from genome-wide association studies. Genet Epidemiol 31:396-407, 2007.

113. Lee S-I, Pe'er D, Dudley A, Church G, Koller D. Identifying regulatory mechanisms using their individual variation reveals key role for chromatin modification. PNAS 103:14062-14067, 2006.

114. Lee S-I, Dudley A, Drubin D, Silver P, Krogan N, Pe'er D, Koller D. Learning a prior on regulatory potential from eQTL data. To appear in PLoS Genetics, 2009.

115. Kudaravalli S, Veyrieras J-B, Stranger BE, DermitzakisET, Pritchard JK. Gene expression levels are a target of recent natural selection in the human genome. Mol Biol Evol 26:649-658, 2009.

116. Tzeng J-Y, Byerley W, Devlin B, Roeder K, Wasserman L. Outlier detection and false discovery rates for whole-genome matching. J Am Statist Assoc 98:236-247, 2003.

117. Tzeng J-Y, Devlin B, Wasserman L, Roeder K. On the identification of disease mutations by haplotype similarity and goodness-of-fit. Am J Hum Genet 72:891-902, 2003.

118. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusis AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich R. Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6:e107, 2008.

119. Ounissi-Benkalha H, Polychronakos C. The molecular genetics of type 1 diabetes: new genes and emerging mechanisms. Trends Mol Med 14:268-275, 2008.

120. Morel PA, Dorman JS, Todd JA, McDevitt HO, Trucco M. Aspartic acid at position 57 of the HLA-DQ beta chain protects against type I diabetes: a family study. PNAS 85:8111-8115, 1988.

121. Dorman JS, LaPorte RE, Stone RA, Trucco M. Worldwide differences in the incidence of type I diabetes are associated with amino acid variation at position 57 of the HLA-DQ beta chain. PNAS 87:7370-7374, 1990.

122. Trucco M, Ball E, Carcassi C, Dombrausky L, Cascino I, Boehm B, Hoover M, +Mehra N, Balakrishnan K, Contu L: RFLP analysis of DQ-beta chain gene: Workshop report. In: *Histocompatibility Testing 1987 Vol.1* p860, 1989.

123. Boehm BO, Trucco M: Immunogenetics and Diabetes Mellitus. International Diabetes Federation Bulletin 34:14, 1989.

124. Boehm B, Manfras B, Rosak C, Kuehnl P, Schoffling K and Trucco M: TCR-alpha and TCR-beta diallelic RFLPs in Insulin-dependent (Type I) Caucasian diabetic patients. Diabetes Research 15:63, 1990.

125. Boehm BO, Manfras B, Seidl S, Holzberger G, Kühnl P, Rosak C, Schöffling K and Trucco M: The HLA-DQß non-Asp 57 allele, a predictor of future Insulin-Dependent Diabetes Mellitus in patients with autoimmune Addison's disease. Tissue Antigens 37:130, 1991.

126. Manfras BJ, Boehm BO, Rudert WA, Thomas HG, Schöffling K and Trucco M, Usadel KH: Primer directed amplification of homologous sequences: specific amplification of CYP21A and CYP21B genes on chromosome six. Advances in Molecular Genetics 4:31, 1991.

127. Boehm B, Manfras B, Rosak C, Schoffling K and Trucco M: Aspartic acid at position 57 of the HLA-DQß chain is protective against future development of Insulin-dependent (type I) diabetes mellitus. Klin Wochenschr 69:146, 1991.

128. Boehm B, Manfras B, Schoffling K, Seissler J, Gluck M, Holzberger G, Seidl S, Kuhnl P, Trucco M and Scherbaum W: Epidemiology and immunogenetic background of islet cell antibody-positive nondiabetic schoolchildren: Ulm-Frankfurt population study. Diabetes 49:1435, 1991.

129. Boehm BO, Seissler J, Gluck M, Manfras B, Thomas H, Schmidt K, Rudert WA, Usadel KH, Trucco M and Scherbaum W: The level and persistence of islet cell antibodies in healthy school-children are associated with polymorphic residues of the HLA-DQß chain. Disease Markers 9:273, 1991.

130. Boehm BO, Scherbaum WA, Schöffling K, Kühnl P, Althoff P, Manfras B, Usadel K-H, Trucco M: Prevalence of HLA-DQ beta chain non-Asp alleles in Type I (insulin-dependent) diabetics with young and older ages on onset. Klin Wochenschr 69:687, 1991.

131. Manfras BJ, Swinyard M, Rudert WA, Ball EJ, Lee PA, Kühnl P, Trucco M, Boehm BO: Altered CYP21 genes in extended HLA-Haplotypes associated with congenital adrenal hyperplasia (CAH) - A family study. Human Genet 92:33, 1993.

132. Manfras B, Rudert WA, Trucco M, Boehm BO: Cloning and characterization of a glutamate transporter cDNA from human brain and pancreas. Biochem Biophys ACTA Journal 1195:185, 1994.

133. Manfras BJ, Rudert WA, Trucco M, Boehm BO: Analysis of the a/b T-cell receptor repertoire by competitive and quantitative family-specific PCR with exogenous standards and high resolution fluorescence based CDR3 size imaging. J Immunological Methods 210:235, 1997.

134. Starzl TE, Demetris AJ, Trucco M, Ramos H, Zeevi A, Rudert WA, Kocova M, Ricordi C, Ildstad S, Murase N. Systemic chimerism in human female recipients of male livers. Lancet 340:876-877, 1992.

135. Starzl TE, Demetris AJ, Trucco M, Ricordi C, Ildstad S, Terasaki PI, Murase N, Kendall RS, Kocova M, Rudert WA, et al. Chimerism after liver transplantation for type IV glycogen storage disease and type 1 Gaucher's disease. N Engl J Med 328:745-749, 1993.

136. Conrad B, Weidmann E, Trucco G, Rudert WA, Behboo R, Ricordi C, Rodriquez-Rilo H, Finegold D, Trucco M. Evidence for superantigen involvement in insulin-dependent diabetes mellitus aetiology. Nature 371:351-355, 1994.

137. Trucco M, Stassi G. Transplantation biology. Educating effector T cells. Nature 380:284-285, 1996.

138. Boehm BO, Manfras B, Seissler J, Schöffling K, Glück M, Holzberger G, Seidl S, Kühnl P, Trucco M, Scherbaum WA. Epidemiology and immunogenetic background of islet cell antibody--positive nondiabetic schoolchildren. Ulm-Frankfurt population study. Diabetes 40:1435-1439, 1991.

139. Ott PA, Herzog BA, Quast S, Hofstetter HH, Boehm BO, Tary-Lehmann M, Durinovic-Bello I, Berner BR, Lehmann PV. Islet-cell antigen-reactive T cells show different expansion rates and Th1/Th2 differentiation in type 1 diabetic patients and healthy controls. Clin Immunol 115:102-114, 2005.

140. Durinovic-Belló I, Jelinek E, Schlosser M, Eiermann T, Boehm BO, Karges W, Marchand L, Polychronakos C. Class III alleles at the insulin VNTR polymorphism are associated with regulatory T-cell responses to proinsulin epitopes in HLA-DR4, DQ8 individuals. Diabetes 54:S18-24, 2005.

141. Barnstable CJ, Bodmer WF, Brown G, Galfre G, Milstein C, Williams AF, Ziegler A. Production of monoclonal antibodies to group A erythrocytes, HLA and other human cell surface antigens-new tools for genetic analysis. Cell 14:9-20, 1978.

142. Klohe EP, Watts R, Bahl M, Alber C, Yu WY, Anderson R, Silver J, Gregersen PK, Karr RW. Analysis of the molecular specificities of anti-class II monoclonal antibodies by using L cell transfectants expressing HLA class II molecules. J Immunol 141:2158-2164, 1988.

143. Ziegler A, Heinig J, Müller C, Götz H, Thinnes FP, Uchańska-Ziegler B, Wernet P. Analysis by sequential immunoprecipitations of the specificities of the monoclonal antibodies TU22,34,35,36,37,39,43,58 and YD1/63.HLK directed against human HLA class II antigens. Immunobiology 171:77-92, 1986.

Goal 2. *Expand the focus of the T1D research program to include next generation sequencing methods.*

HLA NEXT GENERATION PYROSEQUENCE BASED TYPING

Goal 2. SUMMARY

The complex of Human Leukocyte Histocompatibility Antigens (*HLA*) encoding genes represents a highly polymorphic region of the genome with substantial linkage disequilibrium (LD) between loci extending over 4 megabases of Chromosome 6p21.3. The large number of polymorphic residues in these loci result in a high frequency of heterozygous genotypes. Laboratory methods for genotyping *HLA* alleles are frequently confounded due to the large number of heterozygous combinations that result in ambiguous genotypes even when Sanger DNA sequencing is used. Next generation sequencing methods, such as, the GS-FLX pyrosequencer, utilize cloning by polymerase chain reaction (PCR) amplification and are capable of highest resolution genotyping of *HLA* loci. Development and validation of next generation pyrosequencing based methods for genotyping *HLA* loci are critical for achieving routine high resolution typing of this gene system.

Goal 2. BACKGROUND

The biological role of *HLA* encoded proteins is to act within the immune system presenting antigenic peptides to T-lymphocytes (Klein and Sato, 2000). The large number of polymorphic residues within individual *HLA* loci result in over 2,496 class I and 1,032 class II alleles (Robinson et al., 2003). *HLA-B* contains the greatest number of allele variants, with 1,178 currently reported (Robinson et al., 2003). High resolution genotyping of *HLA* loci is a critical element in determining success of solid organ and bone marrow transplantation (Ringquist et al., 2007a). Transplantation genetics generally involves matching as closely as possible multiple *HLA* loci present in recipient with donor, that is, 2 allele matches for *HLA* class I loci *-A*, *-B*, *-C* as well as class II loci *HLA-DRB1*, *-DQB1*, and *-DPB1* (Hurley et al., 2000).

The molecular basis for genetic variability within the *HLA* system is the presence of stably inherited variations of genomic DNA. These polymorphisms are present in the *HLA* coding regions with the greatest number occurring in exon 2 (Ringquist et al., 2007a). The frequency of different *HLA* alleles vary among human geographic populations (Cao et al., 2001; Klitz et al., 2003) probably reflecting ancient human migration patterns as well as more recent admixture events (Cavalli-Sforza et al., 1994; Tu et al., 2007). In any

individual there can be no more than two alleles for any *HLA* locus. However, the large number of allele combinations represented in the human population results from evolutionary pressure to maintain immune surveillance of as great a number of potential pathogen-derived peptides as possible (Moore et al., 2002; Trachtenberg et al., 2003).

Numerous ambiguities have been documented when genotyping *HLA* alleles (Adams et al., 2004) and primarily result from the presence of cis/trans combinations of polymorphic residues occurring in certain heterozygous allele pairings (Robinson et al., 2003). These ambiguities occur even in Sanger based sequencing strategies. One possible remedy for this issue is to clone individual alleles prior to sequencing. However, in the event that many patient samples require genotyping the burden associated with this step is substantial. Next generation pyrosequencing methods represent a convenient approach to *HLA* genotyping. The methodology exploits emulsion PCR as a cloning step performed prior to sequencing (Shendure and Hanlee, 2008). Among the advantages of the approach are that hundreds of thousands of individual PCR steps can be performed in parallel. The system enables experimental design in which amplicon libraries of *HLA* gene exons are amplified from large sample cohorts. The resulting material can be sequenced from products cloned by emulsion PCR resulting in highly accurate and efficient *HLA* genotyping.

Goal 2. OBJECTIVE

The project goal is to exploit next generation pyrosequencing in order to perform allele resolution genotyping of *HLA* class I (*HLA-A*, *-B*, *-C*) and class II (*HLA-DRB1/3/4/5*, *-DQB1*, *-DPB1*) loci. DNA samples will be obtained from the Children's Hospital of Pittsburgh (CHP) Histocompatibility Laboratory and provide a collection of previously *HLA* genotyped samples that can be used to validate the method. The final phase of the project will utilize DNA obtained from patient cohorts with Addison disease, Grave's disease, or Type 1 Diabetes Nephropathy (T1DN) with the goal of identifying *HLA* genotypes associated with these autoimmune syndromes.

Goal 2. SPECIFIC AIMS

Aim 1. Design and validate primers for next generation pyrosequencing of *HLA* class I and class II loci.
Aim 2. Optimize experimental conditions to enable accurate and efficient *HLA* genotyping.
Aim 3. Utilize massive parallel pyrosequencing to genotype the *HLA* region.

Goal 2. RESEARCH DESIGN AND METHODS

GS-FLX Titanium Technology: The Roche GS-FLX platform is a pyrosequence-based instrument first introduced by 454 Life Sciences (Margulies et al., 2005). Advantages of the method include sequencing of PCR generated products so long as the amplicon is flanked by adaptor sequences. Clonal PCR products are generated as part of the approach. This is accomplished via emulsion PCR (Dressman et al., 2003) exploiting a step in which amplicons are captured on the surface of microbeads. Following the emulsion PCR step beads are combined, treated to remove unbound DNAs, and then enriched for amplicon bearing beads, that is, those beads that supported productive DNA amplification during the emulsion PCR step. The last step prior to sequencing is to hybridize a universal sequencing adaptor next to the start of the *HLA* region.

Sequencing steps are performed using pyrosequencing (Ronaghi et al., 1996). This method has been applied previously to *HLA* genotyping (Ringquist et al., 2002; Ramon et al., 2003; Ringquist et al., 2004; Entz et al., 2005; Ringquist et al., 2007b; Lu et al., 2009). However, the GS-FLX system represents an improved pyrosequencing platform in that it is performed within hundreds of thousands of individual picoliter scale reaction wells and is capable of reaching read lengths of 500 nucleotides. Given that the length of the relevant *HLA* exons are no greater than 277 base pairs the method is sufficient to generate complete sequencing of exons 2 and 3 from these loci.

Limitations of the method occur during sequencing of homopolymers. Pyrosequencing of homopolymers (e.g., AAAA or GGGG) result in increased signal intensity per sequencing cycle. As the relationship between signal intensity and number of consecutively incorporated nucleotides is nonlinear interpretation of the data from these regions can result in insertion/deletion errors during de novo sequencing. In the case of *HLA*, however, this type of error is anticipated to be correctable since the most frequently occurring *HLA* alleles have been

described (Robinson et al., 2003). When insertion/deletion errors occur within known homoploymer sequences they will be edited using rules of maximum parsimony to align experimental data with the closest known allele.

The GS-FLX instrument is available at the University of Pittsburgh Genomics Core Laboratory. It generates as many as 1,000,000 reads in parallel at lengths up to 500 base pairs. In order to achieve 50-times coverage of any sequence the experimental design should favor creation of 20,000 unique sequences per run. For the *HLA* system in which 3 class I loci (exons 2 and 3) and 3 class II loci (exon 2) will be sequenced there are a total of 18 exons to be sequenced for each heterozygous individual. At maximum capacity the method will allow greater than 1,000 individuals to be sequenced in a single assay. When smaller scale experiments are required the GS-FLX platform can be subdivided into as little as 1/16 of the full-scale setup (roughly 25,000 to 40,000 reaction wells) and is anticipated to result in complete *HLA* genotyping of at least 27 individuals while maintaining 50-times coverage.

Table 1. Cohorts

Cohort	Number
Experimental Aim 1:	
CHP	12
Experimental Aim 3:	
Addison	120
Grave's	200
T1DN	832

DNA Cohorts: *HLA* typed samples are available from the CHP Histocompatibility Laboratory (Table 1). These samples have been genotyped using hybridization strategies and provide in most cases medium resolution *HLA* typing (Ringquist et al., 2007a). This corresponds to genotypes that have been narrowed down to a few (roughly 1 to 6) possible allelic combinations for each locus. The CHP cohort will be used to develop and validate the next generation pyrosequencing approach to *HLA* typing. Case cohorts will be used during the final phase of the project. These have been obtained from patients with 3 different autoimmune diseases which frequently co-occur in patients with Autoimmune Polyendocrine Syndrome 2 (APS2), that is, Addison Disease (N=120), Grave's Disease (N=200), and T1DN (N=832). All materials are available as purified high molecular weight DNA.

Table 2. HLA locus specific primers.

HLA	IHWG Name	Sequence (5'-3')	Comment
Exon 2 Primers:			
-A/-B	AINT1F	GCGCCKGGASGAGGGT	Forward
-A/-B	INT2R	GGATCTCGGACCCGGAG	Reverse
-C	DLFEX2	GGGTCGGGCGGGTCTCAG	Forward
-C	3'97070b	TCGAGGGTCTGGGCGGGTT	Reverse
-DPB	DPB5.1	AGAGGATTAGATGAGAGTGGCG	Forward
-DPB	DPB3.4	CTCACTCCCGAAACCCGGCC	Reverse
-DQB	Upper	TCCTCGCAGAGGATTTCG	Forward
-DQB	Lower	GGGCGACGACGCTCACCTC	Reverse
-DRB	2DRBAMP-A	CCCCACAGCACGTTTCYTG	Forward
-DRB	2DRBAMP-B	CCGCTGCACTGTGAAGCTCT	Reverse
Exon 3 Primers:			
-A	INT2F	TTACCCGGTTTCATTTTCAG	Forward
-A	AAmp4	GGCCCCTGGTACCCGTGCGCTG	Reverse
-B	INT2F	TTACCCGGTTTCATTTTCAG	Forward
-B	BAmp2	CCATCCCCGGCGACCTATAGGAGATG	Reverse
-C	5' 98003	CTCGACCGGAGAGAGCCC	Forward
-C	CAmp2	GGAGATGGGGAAGGCTCCCCACT	Reverse

Study Design and Data Analysis: Experimental Aim 1 of the project will focus on a pilot study using N=12 DNA samples selected from the CHP Histocompatibility Laboratory cohort (Table 1). These DNAs have already been genotyped at medium resolution and will be chosen to reflect the most frequently occurring *HLA* alleles observed among European-Americans (Cao et al., 2001; Klitz et al., 2003). *HLA* locus specific PCR primers (Table 2) have been selected from the International Histocompatibility Working Group (IHWG) recommended primers for amplification of *HLA* class I (exons 2 and 3) and class II (exon 2) loci. Modification of the IHWG set of primers involved addition of an 8 to 10 nucleotide barcode sequence to the 5' end along with flanking sequences taken from the Roche GS-FLX primer A and B (Figure 1). Modified locus specific primers along with the relevant barcodes have already been designed using error-correcting barcode sequences designed to maximize nucleotide differences among barcodes (Hamady et al., 2008).

Figure 1. HLA Locus Specific Amplification Primers

Generic Forward Primer (5'-3')

Roche Primer B	Spacer	2DRAmp-A
GCCTTGCCAGCCCGCTCAG	*TC*	CCCCACAGCACGTTTCYTG

Generic Reverse Primer (5'-3')

Roche Primer A	Barcode	Spacer	2DRAmp-B
GCCTCCCTCGCGCCATCAG	**NNNNNNNNN**	*CA*	CCGCTGCACTGTGAAGCTCT

There will be 8 PCR reactions performed on each of 12 samples and are expected to amplify class I and class II *HLA* loci. Amplification of *HLA* exons will be confirmed by monitoring the size of each PCR product using 1% agarose gel electrophoresis. The anticipated PCR product sizes are listed in Table 3.

The PCR product from each amplification reaction will be purified using calibrated AMPure SPRI beads (Agencourt). The concentration of purified materials will be measured by the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen). Purified amplicons will be adjusted to a standard concentration of 10 μM in TE pH7 buffer. Pools of PCR products will be generated so that amplicons of similar length are pooled together (Table 3). The final amplicon pools (Pool 1: *-A*, *-B*, *-C* and Pool 2: *-DPB*, *-DQB*, *-DRB*) will each contain 10 μM total DNA in 50 μl. Pools will be delivered to the University of Pittsburgh Genomics Core Laboratory for final processing, emulsion PCR amplification, and sequencing using one region of a 16-region GS-FLX pyrosequencing gasket. Sequencing will be performed using the primer complimentary to the Roche Primer A sequence. The experimental design is anticipated to provide greater than 50-times sequencing coverage. The high level of coverage will enable testing of the number of *HLA* loci amplified as well as the accuracy of genotyping.

Table 3. Size of PCR products.

	Amplicon Size (bp)		
HLA	Exon 2	Exon 3	Pool
-A	457	439	1
-B	457	479	1
-C	480	488	1
-DPB	388	--	2
-DQB	340	---	2
-DRB	324	---	2

The goals of the experimental Aim 2 will be pursued simultaneously with experimental steps performed during Aim 1 in order to identify areas in which the experimental design should be optimized. These are primarily anticipated to include troubleshooting of *HLA* locus specific PCR conditions (e.g., thermal cycling, temperature, times, and number of cycles) as well as the amount of genomic DNA used in each reaction well. Determination of PCR yield will occur when examining amplicon products by agarose gel electrophoresis. Errors are anticipated to arise due to gene duplication occurring within the *HLA* system and may result in amplification of additional *HLA* loci and pseudogenes. These errors can occur even though PCR assays yield the predicted size product. Providing that the efficiency of competing reactions do not suppress amplification of the targeted loci the capacity of the next generation pyrosequencing platform is likely to be sufficient to maintain greater than 50-times coverage. Therefore, as a result of the excess capacity of the experimental design, sequencing of additional *HLA* loci is not anticipated to present a challenge to data analysis.

In the final phase of the project experimental Aim 3 will be designed to scale the process to sequence class I and class II *HLA* loci to include N=1,152 participates at 50-times coverage. The approach will follow that described previously for experimental Aim 1. The principal differences are that DNA will be obtained from the case cohort of Addison Disease, Grave's Disease, and T1DN (Table 1) and the number of error-correcting barcode sequences will be expanded. Greater than 1,500 error-correcting barcode primers are available from Hamady et al. (2008) and a subset of these will be used to generate the final primer set. PCR amplified samples will be processed as previously described by pooling AMPure purified and PicoGreen calibrated products. Next generation pyrosequencing will be performed on the full GS-FLX platform. The experimental design is anticipated to achieve 50-times coverage for sequencing 1,152 samples at 18 exons.

Justification of Sample Size: There are two GS-FLX pyrosequencing experiments proposed for the project. The first experiment will sequence N=12 CHP Histocompatibility Laboratory samples. The numbers of samples used in this phase of the project were chosen for convenience during sample preparation. There will be 8 PCR amplifications performed per sample and will generate a total of 96 reactions. This can be performed in a single 96-well PCR tray. Likewise, confirmation of PCR yield and size of the resulting PCR product can be determined on a single 96-well 1% agarose gel.

The results from sequencing the CHP Histocompatibility Laboratory samples are expected to provide insight into how well the locus specific *HLA* primers amplify predefined loci and how frequently homologous sequences (e.g., gene duplications and peseudogenes) are also amplified. For example, duplication events within the *HLA* region have created the non-classical *HLA* class I loci *-E*, *-F*, and *-G* as well as multiple pseudogenes, such as, *-DRB2/6/7/8/9* and *-DQB2*. Knowledge of the number of loci amplified by each primer pair will be used to design subsequent experiments so that 50-times sequence coverage is maintained. A second benefit of the design of experimental Aim 1 is that the resulting data can be compared to the results obtained from hybridization based medium resolution *HLA* genotyping. There are too few samples used in this phase of the project to estimate accuracy of the method. High correspondence between the methods will be essential to successfully complete this phase of the project.

GS-FLX pyrosequencing performed during experimental Aim 3 is anticipated to employ as many as N=1,152 samples chosen from among the cohorts of Addison Disease, Grave's Disease, and T1DN (Table 1). The numbers of individual samples from each cohort were chosen to completely use the available Addison Disease and Grave's Disease samples while filling out the remainder of the assay with T1DN cases. The final number of samples is chosen based upon convenience of sample handling. For example, the PCR amplification phase will involve 24 x 384-well reaction trays. The CHP research laboratory is equipped with multiple dual head 384-well thermal cycle instruments that will enable efficient preparation of amplified DNAs. Preparation of reaction trays will be performed using the BioMek FX liquid handling robot available at CHP. Analysis of the data will focus upon determining allele frequencies in the case populations. Final analysis of the data will focus on establishing the respective similarities and differences of the various *HLA* alleles observed in the case cohorts as well as to the expected allele frequencies reported for the European-American population (Cao et al., 2001; Klitz et al., 2003).

Goal 2. TIMELINE (but timeless for now)

Aim 1: Design *HLA* class I and class II specific PCR amplification primers; PCR amplification of *HLA* loci using genomic DNA collected from the CHP Histocompatibility Laboratory cohort; Initiate GS-FLX pyrosequencing at the Genomics Core Laboratory; Analyze resulting data for *HLA* genotyping.

Aim 2: Quality control analysis of PCR amplification product specificity and yield via agarose gel electrophoresis; Analyze sequencing results for *HLA* pseudogenes as well as non-classical loci.

Aim 3: Prepare PCR amplification products from Addison Disease, Grave's Disease, and T1DN cohorts; Initiate GS-FLX pyrosequencing at the Genomics Core Laboratory; Analyze resulting data for *HLA* genotyping.

Goal 2. LITERATURE CITED

Adams, S. D., Barracchini, K. C., Chen, D., Robbins, F., Wang, L., Larsen, P., et al. (2004) Ambiguous allele combinations in HLA Class I and Class II sequence-based typing: when precise nucleotide sequencing leads to imprecise allele identification. J Transl Med 2:30-35.

Cao, K., Hollenbach, J., Shi, X., Shi, W., Chopek, M., and Fernandez-Vina, M. A. (2001) Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations. Hum Immunol 62:1009-1030.

Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1994) The History and Geography of Human Genes. Princeton University Press, Princeton, NJ.

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci U S A 100:8817-8822.

Entz P, Toliat MR, Hampe J, Valentonyte R, Jenisch S, Nürnberg P, Nagy M. (2005) New strategies for efficient typing of HLA class-II loci DQB1 and DRB1 by using Pyrosequencing. Tissue Antigens 65:67-80.

Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R. (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods 5:235-237.

Hurley, C. K., Baxter-Lowe, L. A., Begovich, A. B., Fernandez-Vina, M., Noreen, H., Schmeckpeper, B., et al. (2000) The extent of HLA class II allele level disparity in unrelated bone marrow transplantation: analysis of 1259 National Marrow Donor Program donor-recipient pairs. Bone Marrow Transplant. 25:385-393.

Klein J, Sato A. (2000) The HLA system. First of two parts. N Engl J Med 343:702-709.

Klitz, W., Maiers, M., Spellman, S., Baxter-Lowe, L. A., Schmeckpeper, B., Williams, T. M., and Fernandez-Vina, M. (2003) New HLA haplotype frequency reference standards: high-resolution and large sample typing of HLA DR-DQ haplotypes in a sample of European Americans. Tissue Antigens 62:296-307.

Lu Y, Boehm J, Nichol L, Trucco M, Ringquist S. (2009) Multiplex HLA-typing by pyrosequencing. Methods Mol Biol 496:89-114.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Moore, C. B, John, M., James, I. R, Christiansen, F. T, Witt, C. S, and Mallal, S. A. (2002) Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science 296:1439-1443.

Ramon D, Braden M, Adams S, Marincola FM, Wang L. (2003) Pyrosequencing trade mark : A one-step method for high resolution HLA typing. J Transl Med 1:9.

Ringquist S, Alexander AM, Rudert WA, Styche A, Trucco M. (2002) Pyrosequence-based typing of alleles of the HLA-DQB1 gene. Biotechniques 33:166-175.

Ringquist S, Alexander AM, Styche A, Pecoraro C, Rudert WA, Trucco M. (2004) HLA class II DRB high resolution genotyping by pyrosequencing: comparison of group specific PCR and pyrosequencing primers. Hum Immunol 65:163-174.

Ringquist, S., Nichol, L., and Trucco, M. (2007a) Transplantation genetics, in Emery and Rimoin's Principles and Practice of Medical Genetics 5th Edition (Rimoin, D. L., Connor, J. M., Pyeritz, R. E., and Korf, B. R., eds.), Churchill Livingstone, Philadelphia, PA, pp. 983-1010.

Ringquist S, Styche A, Rudert WA, Trucco M. (2007b) Pyrosequencing-based strategies for improved allele typing of human leukocyte antigen loci. Methods Mol Biol 373:115-134.

Robinson, J., Waller, M. J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L. J., et al. (2003) IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucl Acids Res 31:311-314.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242:84-89.

Shendure J, Ji H. (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135-1145.

Trachtenberg, E., Korber, B., Sollars, C., Kepler, T. B., Hraber, P. T., Hayes, E., et al. (2003) Advantage of rare HLA supertype in HIV disease progression. Nat Med 9:928-935.

Tu B, Mack SJ, Lazaro A, Lancaster A, Thomson G, Cao K, Chen M, Ling G, Hartzman R, Ng J, Hurley CK. (2007) HLA-A, -B, -C, -DRB1 allele and haplotype frequencies in an African American population. Tissue Antigens 69:73-85.

Statement of Plans for the Upcoming Research Period

Goal 1. Finalize SNPs for fine mapping T1D risk elements. **Milestone 1A.** Working with publicly available linkage disequilibrium data define SNP clusters and tag-SNPs for fine mapping of genomic regions associated with T1D.

Goal 2. Fine map genomic regions for association with T1D by genotyping select SNPs. **Milestone 2A.** Prepare DNA samples for laboratory analysis. **Milestone 2B.** Initiate genotyping of DNA samples.

In the fourth and final quarterly scientific progress report (06/01/09 - 08/31/09), we reported on our cumulative results for year 01.

During the recently completed research quarter there were 2 goals. These were to finalize SNPs for fine mapping T1D risk elements and to initiate genotyping of T1D cases and non-T1D control samples. Goal 1 has been completed and the work to accomplish Goal 2 will be performed during the requested no cost extension.

Goal 1. *Finalize SNPs for fine mapping T1D risk elements.* ***Milestone 1A****. Working with publicly available linkage disequilibrium data define SNP clusters and tag-SNPs for fine mapping of genomic regions associated with T1D.*

Goal 1 has been completed. We have worked extensively with data made available by the Wellcome Trust Case Control Consortium (WTCCC, 2007). This dataset contains a genetic study of 1963 cases with T1D and 2938 controls obtained from people living in Great Britain who self-identified as white Europeans. The samples were genotyped with the GeneChip 500K Mapping Array Set (Affymetic Chip). At least 56 regions in the genome have strong statistical support in the literature for association with T1D (Table 1) with reported minor allele frequencies (MAF) exceeding 5% and significant odds ratios (OR) for disease risk (Barrett et al., 2009).

Table 1. SNPs associated with T1D.

dbSNP_ID	Chr	Location	Gene	MAF	OR(95%CI)
rs2269241	1p31.3	63,881,359	*PGM1*	0.19	1.10(1.02-1.18)
rs2476601	1p13.2	114,179,091	*PTPN22*	0.09	1.98(1.82-2.15)
rs2816316	1q31.2	190,803,436	*RGS1*	0.22	
rs3024505	1q32.1	205,006,527	*IL10*	0.17	0.84(0.77-0.91)
rs1534422	2p25.1	12,558,192		0.46	1.08(1.02-1.15)
rs917997	2q12.1	102,437,000	*IL18RAP*	0.23	
rs1990760	2q24.2	162,832,297	*IFIH1*	0.60	1.18(1.11-1.23)
rs3087243	2q33.2	204,477,164	*CTLA4*	0.46	
rs11711054	3p21.31	46,320,615	*CCR5*	0.32	
rs10517086	4p15.2	25,694,609		0.30	1.09(1.02-1.17)
rs4505848	4q27	123,351,942	*IL2*	0.34	
rs6897932	5p13.2	35,910,332	*IL7R*	0.71	1.12(1.06-1.19)
rs9268645	6p21.32	32,504,030	*HLA*	0.38	
rs11755527	6q15	91,014,952	*BACH2*	0.47	1.13(1.08-1.19)
rs9388489	6q22.32	126,740,412	*C6orf173*	0.45	1.17(1.10-1.24)
rs2327832	6q23.3	138,014,761	*TNFAIP3*	0.18	
rs1738074	6q25.3	159,385,965	*TAGAP*	0.49	
rs7804356	7p15.2	26,858,190		0.24	0.88(0.82-0.94)
rs4948088	7p12.1	50,994,688	*COBL*	0.05	0.77(0.67-0.90)
rs7020673	9p24.2	4,281,747	*GLIS3*	0.50	0.88(0.83-0.93)
rs12251307	10p15.1	6,163,501	*IL2RA*	0.09	
rs11258747	10p15.1	6,512,897	*PRKCQ*	0.24	
rs10509540	10q23.31	90,013,013	*C10orf59*	0.29	0.75(0.70-0.80)
rs7111341	11p15.5	2,169,742	*INS*	0.32	
rs4963879	12p13.31	25,430,152	*CD69*	0.37	1.09(1.02-1.16)
rs2292239	12q13.2	54,768,447	*ERBB3*	0.34	1.28(1.21-1.35)
rs1678536	12q13.3	56,265,457	*Multiple*	0.31	
rs3184504	12q24.12	110,368,991	*SH2B3*	0.41	
rs1465788	14q24.1	68,333,352		0.29	0.86(0.80-0.91)
rs4900384	14q32.2	97,568,704		0.29	1.09(1.02-1.16)
rs3825932	15q25.1	77,022,501	*CTSH*	0.32	1.16(1.10-1.22)

rs12708716	16p13.13	11,087,374	*CLEC16A*	0.68	1.23(1.16-1.30)
rs12444268	16p12.3	20,250,073		0.30	1.10(1.03-1.17)
rs4788084	16p11.2	28,447,349	*IL27*	0.42	0.86(0.81-0.91)
rs7202877	16q23.1	73,804,746		0.10	1.28(1.17-1.41)
rs16956936	17p13.1	7,574,417		0.14	0.92(0.84-1.00)
rs2290400	17q12	35,319,766	*ORMDL3*	0.50	0.87(0.82-0.93)
rs7221109	17q21.2	36,023,812		0.35	0.95(0.89-1.01)
rs1893217	18p11.21	12,799,340	*PTPN2*	0.19	
rs763361	18q22.2	65,682,622	*CD226*	0.47	1.16(1.10-1.22)
rs425105	19q13.32	51,900,321		0.16	0.86(0.79-0.93)
rs2281808	20p13	1,558,551		0.36	0.90(0.84-0.95)
rs11203203	21q22.3	42,709,255	*UBASH3A*	0.28	
rs5753037	22q12.2	28,911,722		0.39	1.10(1.04-1.17)
rs229541	22q13.1	35,921,264	*C1QTNF6*	0.43	1.04(0.97-1.12)
rs2664170	Xq28	153,598,796		0.32	1.16(1.07-1.24)

Data summarized from Barrett et al. (2009).

Working closely with our collaborators Bernie Devlin (University of Pittsburgh) and Kathryn Roeder (Carnegie Mellon University) we reanalyzed these data. Of the 469,612 SNPs passing quality control performed by the WTCCC, we removed an additional 594 SNPs with poor genotype clustering patterns and all SNPs on Chromosome X. Next we restricted the analysis to those with p-values less than 0.017. From the remaining 10,000 SNPs, we chose SNPs using H-clust, set to pick tag SNPs with squared correlation (r^2) less than 0.04 and MAF greater than 0.01; for a cluster of SNPs in linkage disequilibrium (LD), H-clust used preference scores based on the univariate p-values for association of each SNP (Rinaldo et al., 2005). In this way, our tag SNP selection process included the SNP most likely to be associated with T1D within each LD block. After this process, we further ensured that the set of tag SNPs included no SNPs with r^2 greater than 0.045 on the same chromosome. The resulting set of tag SNPs included 3437 SNPs. We recoded the genotype data for the additive model of inheritance.

Our results are similar to the WTCCC's univariate analysis (Table 2). All 5 of their best signals appeared as significant effects in our model. In addition, on Chromosome 4, SNP rs17388568, which was borderline significant (5.7×10^{-7}) using conditional logistic regression, is also borderline in our analysis (multiple testing corrected p-value=0.35). Our model also identified 4 additional SNPs in the HLA region. Because we restricted our analysis to tag SNPs the LD between these SNPs is minor, suggesting the signal in the MHC is due to multiple variants.

Table 2. Univariate analysis of publicly available T1D data.

Chr	Location	dbSNP_ID	CHP[1] (p-value)	WTCCC[2] (p-value)	Gene
1	114,105,331	rs6679677	5.6×10^{-14}	5.1×10^{-25}	*PTPN22*
4	123,548,812	rs17388568	0.35	5.7×10^{-7}	*IL2*
6	31,735,428	rs2242655	1.8×10^{-2}	5.4×10^{-6}	*HLA-B*
6	32,297,010	rs415929	1.8×10^{-2}	2.9×10^{-5}	*NOTCH4*
6	32,712,350	rs9272346	1.1×10^{-76}	8.9×10^{-122}	*HLA-DR, -DQ*
6	32,910,181	rs241432	3.5×10^{-4}	1.7×10^{-6}	*TAP2*
6	33,111,665	rs448733	2.7×10^{-2}	1.1×10^{-5}	*HLA-DPB*
12	54,756,892	rs11171739	2.6×10^{-5}	1.3×10^{-11}	*ERBB3*
12	110,971,201	rs17696736	1.3×10^{-2}	1.0×10^{-11}	*C12orf30*
16	11,115,395	rs9746695	1.4×10^{-3}	9.6×10^{-9}	*CLEC16A*

1. Best SNP main effects found via CHP analysis are corrected for multiple testing.
2. Nominal univariate p-values reported by WTCCC are not corrected for multiple testing and were obtained using logistic regression.

Applying analysis methods developed during the research project we identified three SNP-SNP interactions as significant (Table 3), corresponding to univariate SNP-SNP p-values that would not have been sufficient to attain genome-wide significance in a standard analysis. This suggests that the new analysis procedure can indeed be more powerful than a series of univariate tests, especially when searching through the vast model space of SNP-SNP interactions.

Table 3. Analysis of paired SNPs associated with T1D risk.

Chr	Location	dbSNP_ID	p-value	r^2	Distance	Genes
cis-acting pair 1:						
6	32,911,818	rs241429	3.5×10^{-66}	0.02	1,637	*TAP2*
6	32,910,181	rs241432				
cis-acting pair 2:						
6	32,911,818	rs241429	1.9×10^{-25}	<0.001	199,468	*HLA-DR, -DQ*
6	32,712,350	rs9272346				*TAP2*
cis-acting pair 3:						
12	110,918,887	rs11066119	2.3×10^{-15}	<0.001	52,314	*C12orf30*
12	110,971,201	rs17696736				

Best SNP interaction effects found via CHP analysis of paired SNPs and corrected for multiple testing.

Two of the pairs involve SNPs in the MHC region (Table 3). Both of these include a SNP that was identified as a main effects paired with another that did not demonstrate significant main effects (rs241429, univariate p-value of 8.2×10^{-5}) paired with SNPs rs241432 (p-value 1.7×10^{-6}) as well as rs9272346 (p-value 9.0×10^{-122}). The remaining interaction involves a pair of SNPs on Chromosome 12, one discovered as a main effect (rs17696736, p-value 1.0×10^{-11}) that tags the *SH2B3/PTPN11* region, paired with (rs11066119, univariate p-value of 9.6×10^{-5}). Pairs of SNPs are not in linkage disequilibrium (Table 3). Moreover, because each of these variants is significant in the multivariate model, we can conclude that each variant exhibits a significant association with the phenotype, after accounting for all of the other variants in the model. The genotype-by-genotype counts support our findings (Table 4).

Table 4. Genotype-by-genotype counts for paired SNPs.

SNP Pair rs241432 and rs241429:

		rs241432		
		0	1	2
rs241429	0	691	439	250
		(506)	(649)	(223)
	1	707	1562	156
		(890)	(1143)	(392)
	2	401	311	386
		(403)	(518)	(177)

SNP Pair rs9272346 and rs241429:

		rs9272346		
		0	1	2
rs241429	0	751	462	165
		(704)	(542)	(131)
	1	1226	986	213
		(1239)	(954)	(231)
	2	528	481	89
		(561)	(432)	(105)

SNP Pair rs17696736 and rs11066119:

	rs17696736		
	0	1	2

rs11066119	0	1173	2052	865
		(1218)	(2043)	(830)
	1	264	392	122
		(232)	(389)	(158)
	2	22	4	7
		(10)	(16)	(7)

Genotype-by-genotype counts (expected counts) for those SNP-SNP pairs with significant interactions.

Application of the method to data obtained from the Wellcome Trust Case Control Consortium study of Type 1 Diabetes cases and controls uncovered evidence supporting multiple *HLA* class II independent T1D associations within the HLA class I regions occurring at *HLA-B* and *HLA-A* (Valdes et al., 2005; Nejentsev et al., 2007; Howson et al., 2009). Analyses of interacting SNP pairs discovered association occurring within HLA class II as well as within the Chromosome 12q24 region. The HLA region represents the largest genetic risk element for T1D as well as other autoimmune diseases (Klein and Sato, 2000). A likely mechanism by which certain *HLA* alleles influence T1D susceptibility is related to their ability to bind and present autoantigens to autoreactive T-lymphocytes in the thymus (Todd et al., 1987; Morel et al., 1988; Nepom and Erlich, 1991). Likewise, the Chromosome 12q24 region has been confirmed as associated with T1D (Todd et al., 2007; Barrett et al., 2009). These studies have identified a large LD block, estimated at greater than 1.2Mb, harboring at least 2 genes with possible functional relevance to T1D, such as *PTPN11* and *SH2B3* (Todd et al., 2007; Smyth et al., 2008).

Goal 2*. Fine map genomic regions for association with T1D by genotyping select SNPs.* ***Milestone 2A****. Prepare DNA samples for laboratory analysis.* ***Milestone 2B****. Initiate genotyping of DNA samples.*

Milestone 2A, preparation of DNA samples for laboratory analysis, has been completed. DNA samples collected from N=2942 of T1D cases and N=2750 of non-T1D controls have been formatted onto 96-well reaction trays. These samples have been adjusted to a concentration of 10 ng/ul and are available for SNP typing during the requested no cost extension.

Milestone 2B will be completed during the requested no cost extension. The goal is to evaluate regions of the genome associated with T1D risk elements. The research plan is to begin by sequencing select HLA exons and will use the pyrosequencing-based platform developed by 454 Life Sciences. The method is capable of sequencing as many as 500bp and will easily sequence HLA exons of less than 400bp. The research plan is to initiate sequencing of HLA exons using DNA provided by N=8 human subjects. These samples have been previously genotyped by the Children's Hospital of Pittsburgh Histocompatibility Typing Laboratory and will provide a proof of principle for the approach (Table 5). We anticipate that in the sequencing experiment the data will confirm the genotype and enable rationale modification of the protocol to increase the efficiency of the method.

Table 5. HLA genotypes for N=8 samples available for exon sequencing.

HLA Class I Loci:

Subject	*HLA-A*	*HLA-B*	*HLA-Cw*
ACK100	*0201g, *2402g	*2705g, *4402g	*0102, *0501g
ACK101	*0201g, *0201g	*1801g, *4001g	*0304, *0701g
ACK102	*0201g, *0301g	*1501g, *1801g	*0304, *0701g
ACK103	*0201g, *2902g	*1801g, *4404	*0701g, *1601
ACK104	*0201g, *0201g	*1801g, *4001g	*0304, *0701g
ACK106	*0201g, *2402g	*1801g, *4402g	*0501g, *0701g
ACK107	*0201g, *2501	*1801g, *4001g	*0304, *1203
ACK108	*0201g, *6801g	*2703, *5701	*0202, *0701g

HLA Class II Loci:

Subject	*HLA-DRB1*	*HLA-DQB1*
ACK100	*0401, *1302	*0301g/*0302, *0604
ACK101	*0404, *1104	*0301g, *0302
ACK102	*0407, *1104	*0301g, *0301g
ACK103	*0403, *1101	*0301g, *0302
ACK104	*0401, *0801	*0301g/*0302, *0402
ACK106	*0801, *1104	*0301g, *0402
ACK107	*0101, *1301	*0501, *0603
ACK108	*0101, *0701	*0303, *0501

Literature Cited:

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Plagnol V, Pociot F, Schuilenburg H, Smyth DJ, Stevens H, Todd JA, Walker NM, Rich SS; The Type 1 Diabetes Genetics Consortium. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41:703-707.

Howson JM, Walker NM, Clayton D, Todd JA; Diabetes Genetics Consortium. (2009) Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A. Diabetes Obes Metab 11(Suppl 1):31-45.

Klein J, Sato A. (2000) The HLA system. Second of two parts. N Engl J Med 343:782-786.

Morel PA, Dorman JS, Todd JA, McDevitt HO, Trucco M. (1989) Aspartic acid at position 57 of the HLA-DQ beta chain protects against type I diabetes: a family study. Proc Natl Acad Sci USA 85:8111-8115.

Nejentsev S, Howson JM, Walker NM, Szeszko J, Field SF, Stevens HE, Reynolds P, Hardy M, King E, Masters J, Hulme J, Maier LM, Smyth D, Bailey R, Cooper JD, Ribas G, Campbell RD, Clayton DG, Todd JA; Wellcome Trust Case Control Consortium. (2007) Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. Nature 450:887-892.

Nepom GT, Erlich H. (1991) MHC class-II molecules and autoimmunity. Annu Rev Immunol 9:493-525.

Rinaldo A, Bacanu SA, Devlin B, Sonpar V, Wasserman L, Roeder K. (2005) Characterization of multilocus linkage disequilibrium. Genet Epidemiol 28:193-206.

Smyth DJ, Plagnol V, Walker NM, Cooper JD, Downes K, Yang JH, Howson JM, Stevens H, McManus R, Wijmenga C, Heap GA, Dubois PC, Clayton DG, Hunt KA, van Heel DA, Todd JA. (2008) Shared and distinct genetic variants in type 1 diabetes and celiac disease. N Engl J Med 359:2767-2777.

Todd JA, Bell JI, McDevitt HO. (1987) HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. Nature 329:599-604.

Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, Lowe CE, Szeszko JS, Hafler JP, Zeitels L, Yang JH, Vella A, Nutland S, Stevens HE, Schuilenburg H, Coleman G, Maisuria M, Meadows W, Smink LJ, Healy B, Burren OS, Lam AA, Ovington NR, Allen J, Adlem E, Leung HT, Wallace C, Howson JM, Guja C, Ionescu-Tîrgovişte C; Genetics of Type 1 Diabetes in Finland, Simmonds MJ, Heward JM, Gough SC; Wellcome Trust Case Control Consortium, Dunger DB, Wicker LS, Clayton DG. (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39:857-864.

Valdes AM, Erlich HA, Noble JA. (2005) Human leukocyte antigen class I B and C loci contribute to Type 1 Diabetes (T1D) susceptibility and age at T1D onset. Hum Immunol 66:301-313.

Wellcome Trust Case Control Consortium (WTCCC). (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678.

KEY RESEARCH ACCOMPLISHMENTS:

1. Creation of a sample repository containing DNA from N=2942 T1D and N=2750 non-T1D participants.

2. Establishment of laboratory methods for multiplex SNP genotyping.

3. Created advanced statistical methods for analysis of SNPs and SNP-SNP pairs for association with T1D

4. Initiation of T1D genetic studies to evaluate candidate genetic markers.

5. Analysis of genetic data to evaluate candidate genes for association with diabetes phenotypes.

6. Development of laboratory protocols for multiplex sequencing of genes associated with T1D.

7. Publication of 10 manuscripts.

REPORTABLE OUTCOMES:

1. Zhang, L., Ringquist, S., Perdomo, G., Qu, S., Trucco, M., and Dong, H.H. Proteomic analysis of fructose-induced fatty liver in hamsters. Metabolism 57, 1115-1124 (2008).

2. Lu, L., Boehm, J., Nichol, L., Trucco, M., and Ringquist, S. Multiplex HLA typing by pyrosequencing. In Methods in Molecular Biology, vol 496: DNA and RNA Profiling in Human Blood. ed. P. Bugert. Humana Press Inc., Totowa, New Jersey (2009).

3. Wu J, Devlin B, Ringquist S, Trucco M, and Roeder K. Screen and clean: a tool for identifying interactions in genome-wide association studies. Genetic Epidemiology (submitted).

Abstracts; Presentations; Patents and Licenses Applied for and/or Issued; Degrees Obtained that are Supported by this Award: None

Development of Cell Lines, Tissue or Serum Repositories
1. Repository of DNA samples collected from T1D patient cases and non-T1D controls exceeding 5,690 subjects.

Informatics such as Databases and Animal Models, etc: None

Funding Applied for Based on Work Supported by this Award
1. An application for funding has been submitted to the National Institutes of Health in response to RFA-DK-08-006 Fine Mapping and Function of Genes for Type 1 Diabetes. The project is entitled "Refining the Genetic and Functional Architecture of Type 1 Diabetes" and the goal is to elucidate molecular networks affecting T1D susceptibility that are directly influenced by stably inherited genetic variants. Increased understanding of the gene-networks underlying disease risk will aid in the development of accurate screening tools as well as in creation of new therapeutic treatments.

Employment or Research Opportunities Applied for and/or Received Based on Experience/Training Supported by this Award: None

CONCLUSION:

The conclusions for the first and second year of funding are that statistically significant genetic markers associated with T1D have been identified using novel statistical methods for analysis of genome-wide association studies. Statistically significant signals occurring on Chromosomes 6p and 12q correspond with paired SNP-SNP interactions. Work that is planned for the upcoming year (covered by the requested no-cost extension) will seek to evaluate the regions of interacting SNPs discovered during this study in order to evaluate their association with causal models for altered risk of developing disease.

The research project generated 10 publications. These are listed under the section entitled "REPORTABLE OUTCOMES".

The So What Section. What are the implications of this research? Diabetes affects 16 million Americans and 800,000 new cases annually. African, Hispanic, Native and Asian Americans are particularly susceptible to its most severe complications. Costs associated with diabetes may be as high as $132 billion. Diabetes accounts for 42% of new cases of end-stage renal disease with over new 100,000 cases per year at an average cost of $55,000 per patient annually.

What are the military significance and public purpose of this research? As the military is a reflection of the U.S. population improved prediction of risk for developing diabetes and diabetic complications among active duty members of the military, their families, and retired military personnel will potentially allow focused preventative treatment of at risk individuals, providing significant healthcare savings and improved patient well being.

REFERENCES:

1. Ringquist, S, Styche, A., Rudert, W.A., and Trucco, M. Pyrosequence-based strategies for improved allele typing of HLA loci. In: Methods in Molecular Biology, vol 373: Pyrosequencing Protocols. ed. S. March. Humana Press Inc., Totowa, New Jersey (2007).
2. Ringquist, S., Pecoraro, C., Lu, Y., Styche, A., Rudert, W.A., Benos, P.V., and Trucco, M. Web-based primer design software for genome scale SNP mapping by pyrosequencing. In: Methods in Molecular Biology, vol 373: Pyrosequencing Protocols. ed. S. March. Humana Press Inc., Totowa, New Jersey (2007).
3. Ringquist, S., Nichol, L., and Trucco, M. Transplantation Genetics. In Emery and Rimoin's Principles and Practice of Medical Genetics. eds. D.L. Rimoin, M. Conner, R.E. Pyeritz, B.R. Korf, and A.E. Emery 5th Edition, Elsevier Books, Oxford (2007).
4. Pasquali, L., Bhargava, R., Ringquist, S., Styche, A., Bedeir, A., and Trucco, G. Quantitative methylation of CpG islands in progressive breast neoplastic lesions from normal to invasive carcinoma. Cancer Letters 257, 136-144 (2007).
5. Ringquist, S., Pecoraro, C., and Trucco, M. Web-based program for pyrosequencing primer design. ASHI Quarterly 31, 50-52 (2007).
6. Pasquali, L., Trucco, M., and Ringquist, S. Navigating pathways affecting type 1 diabetic kidney disease. Pediatric Diabetes 8, 307-322 (2007).
7. Luca, D., Ringquist, S., Klei, L., Lee, A.B., Gieger, C., Wichmann, H.-E., Schreiber, S., Krawczak, M., Lu, Y., Styche, A., Devlin, B., Roeder, K., and Trucco, M. On the use of general control samples for genome-wide association scans: genetic matching highlights causal variants. American Journal of Human Genetics 82, 452-463 (2008).
8. Zhang, L., Ringquist, S., Perdomo, G., Qu, S., Trucco, M., and Dong, H.H. Proteomic analysis of fructose-induced fatty liver in hamsters. Metabolism 57, 1115-1124 (2008).
9. Lu, L., Boehm, J., Nichol, L., Trucco, M., and Ringquist, S. Multiplex HLA typing by pyrosequencing. In Methods in Molecular Biology, vol 496: DNA and RNA Profiling in Human Blood. ed. P. Bugert. Humana Press Inc., Totowa, New Jersey (2009).
10. Wu J, Devlin B, Ringquist S, Trucco M, and Roeder K. Screen and clean: a tool for identifying interactions in genome-wide association studies. Genetic Epidemiology (submitted).

In the first quarterly scientific progress report of year 02 of our project (09/01/09 - 11/31/09), we reported on the following:

Type 1 Diabetes Genome-Wide Association and Fine Mapping Study Update:

Our efforts during the recently completed research quarter have focused on developing laboratory protocols for high throughput *HLA* genotyping of Type 1 Diabetes (T1D) cases and non-T1D control samples using next-generation DNA sequencing methods. Improved *HLA* genotyping can be used to increase genetic screening of at-risk individuals and to investigate the possible connection between select *HLA* alleles (e.g., *HLA-DRB1*04*) and risk of developing diabetes complications (Cordovado et al., 2008).

Table 1. Cohort 1 HLA Genotypes.

Subject	*-A*	*-B*	*-Cw*	*-DPB1*	*-DQB1*	*-DRB1*
ACK100	*02G1	*270502	*010201	*010101	*03G1	*040101
	*24G1	*44G1	*050102	*020102	*060401	*130201
ACK101	*02G1	*18G1	*030401	*030101	*03G1	*0404
	*02G1	*40G1	*07G1	*040101	*030201	*110401
ACK102	*02G1	*15G1	*030401	*0402	*03G1	*040701
	*03G1	*18G1	*07G1	*050101	*03G1	*110401
ACK103	*02G1	*18G1	*07G1	*0601	*03G1	*040301
	*290201	*4404	*160101	*1401	*030201	*1101G1
ACK104	*02G1	*18G1	*030401	*010101	*03G1	*040101
	*02G1	*40G1	*07G1	*040101	*0402	*080101
ACK106	*02G1	*18G1	*050101	*020102	*03G1	*080101
	*24G1	*44G1	*07G1	*030101	*0402	*110401
ACK107	*02G1	*18G1	*030401	*0402	*0501	*010101
	*250101	*40G1	*120301	*1401	*060301	*130101
ACK108	*02G1	*2703	*020201	*050101	*030302	*010101
	*680101	*570101	*07G1	*0601	*050101	*07010101

Genotypes for *HLA-A*, *-B*, *-Cw*, and *-DRB1* were inferred using the oligo hybridization and the allele frequency data. *HLA-DQB1* genotypes were inferred from haplotype frequency tables. *HLA-DPB1* genotypes are selected from among the most frequently observed alleles.

Goal 1 of the recently completed Research Quarter: The goal of the project has been to evaluate regions of the genome associated with T1D risk elements. Specific HLA haplotypes represent the strongest genetic risk element for developing T1D. For example, the *HLA-DRB1*0405-DQA1*0301-DQB1*0302* and *HLA-DRB1*0301-DQA1*0501-DQB1*0201* haplotypes are associated with increased risk and exhibit Odds Ratios (OR) of OR=11.4 (95%CI, 2.7-47.7) and OR=3.64 (95%CI, 2.9-4.6), respectively (Erlich et al., 2008). The experimental plan is to use the GS-FLX pyrosequencing-based platform developed by Roche Diagnostics Corporation to sequence select *HLA* exons from *HLA* class I loci *-A*, *-B*, *-C* and class II loci *-DPB1*, *-DQB1*, *-DRB1*. Accurate methods for genotyping these loci are important for developing methods for predicting risk to autoimmune diseases such as T1D and diabetes complications as well as for use in clinical settings involving organ transplantation. The GS-FLX next generation sequencing method is capable of sequencing read lengths of roughly 250bp and will easily sequence HLA exons when sequencing is performed on both the forward and reserve DNA strands. A notable advantage of the instrument is that it is capable of sequencing multiple samples in parallel, analyzing up to 400,000 clones in each run (Shendure and Ji, 2008). Working with the standard protocol that enables each sequence to be analyzed with at least 50-times coverage the method, when working at maximal capacity, will enable analysis of 8,000 individual exons from 4,000 heterozygous subjects. DNA has already been prepared from *HLA-A, -B, -C* exons 2 and 3 as well as *HLA-DPB1, -DQB1, -DRB1* exon 2 using material collected from N=8 human subjects (Table 1). These samples have been previously genotyped by the Children's Hospital of Pittsburgh (CHP) Histocompatibility Typing Laboratory and will provide proof of principle for the approach. We anticipate that in the sequencing experiment the data will confirm the genotype and will enable rationale modification of the protocol to optimize efficiency of the method.

Progress meeting the goal. Design and validation of PCR amplification protocols have been completed. Amplification primers are shown in Table 2 and their annealing positions relative to the targeted *HLA* exons are illustrated in Figure 1. In order to use the primers in the GS-FLX based next generation sequencing method the amplification primers are modified with a 5' fixed sequence Roche Primer A (5'-GCCTCCCTCGCGCCATCAG-3') and Roche Primer B (5'-GCCTTGCCAGCCCGCTCAG-3') to support emulsion PCR (Figure 1, Upper Panel). Primers are also modified with an internal barcode sequence (e.g., 5'-ACGAGTGCGT-3') to identify the subject from whom the amplified material originated. The location of each primer relative to the selected exon that has been targeted for sequencing is illustrated in the lower panel (Figure 1). For example, the forward and reverse primers for *HLA* loci *-A*, *-B*, *-C*, and *-DPB* are located in the introns flanking each exon. In contrast, the forward primer used to amplify *HLA-DQB* exon 2 anneals to the intron/exon boundary while the reverse primer is located in intron 3. For *HLA-DRB* the forward and reverse amplification primers anneal at opposite ends of the exon 2 sequence and overlap the intron/exon boundary (Figure 1, Lower Panel).

Table 2. PCR Amplification Primers.

Locus	Exon	Comment	Tm	Sequence
-A/-B	2	Forward	63C	GCGCC**K**GGA**S**GAGGGT
		Reverse	57C	GGATCTCGGACCCGGAG
-C	2	Forward	64C	GGGTCGGGCGGGTCTCAG
		Reverse	64C	TCGAGGGTCTGGGCGGGTT
-DPB	2	Forward	56C	AGAGGATTAGATGAGAGTGGCG
		Reverse	64C	CCGGCCCAAAGCCCTCACTC
-DQB	2	Forward	54C	TCCTCGCAGAGGATTTCG
		Reverse	63C	GGGCGACGACGCTCACCTC
-DRB	2	Forward	58C	CCCCACAGCACGTTTC **Y**TG
		Reverse	60C	CCGCTGCACTGTGAAGCTCT
-A	3	Forward	51C	TTACCCGGTTTCATTTTCAG
		Reverse	69C	GGCCCCTGGTACCCGTGCGCTG
-B	3	Forward	51C	TTACCCGGTTTCATTTTCAG
		Reverse	61C	CCCACTGCCCCTGGTACC
-C	3	Forward	60C	CTCGACCGGAGAGAGCCC
		Reverse	65C	GGAGATGGGGAAGGCTCCCCACT

Figure 1. Locus Specific Amplification Primers.

Barcoded Amplification Primers

Roche Primer B	Barcode	Spacer	Forward Primer *HLA-DQB*
GCCTTGCCAGCCCGCTCAG	**ACGAGTGCGT**	*TC*	TCCTCGCAGAGGATTTCG

Roche Primer A	Barcode	Spacer	Reverse Primer *HLA-DQB*
GCCTCCCTCGCGCCATCAG	**ACGAGTGCGT**	*CA*	GGGCGACGACGCTCACCTC

Schematic PCR Design

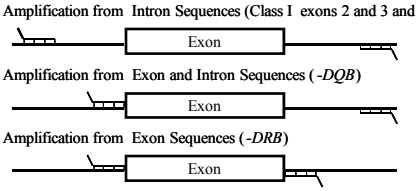
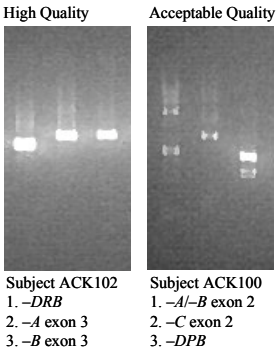


Figure 2. Analysis of DNA Amplicons



Samples collected from N=8 subjects have been prepared as DNA amplicons of 9 exons from 6 HLA loci (*HLA-A, -B, -C* exons 2 and 3 and *HLA-DRB, -DQB, -DPB* exon 2). Figure 2 summarizes the quality of amplified DNA generated for use in next generation sequencing assays. The left hand panel illustrates typical results obtained from the highest quality amplicons that were generated. The PCR generated DNAs are of the expected size for *-DRB* exon 2 (355 bp), *-A* exon 3 (456bp) and *-B* exon 3 (453bp). In contrast, the right hand panel shows the results of amplicon DNA that was of lesser quality. These amplification reactions did, however, generate bands of the expected size (i.e., *-A/-B* exon 2, 477bp; *-C* exon 2, 492bp, and *-DPB* exon 2, 400bp) although additional PCR products were also generated. In the experiment that is currently underway material represented by the left hand panel will be analyzed in order to evaluate the sequencing results when only the highest quality material is used. In a parallel sequencing assay material represented by both panels (i.e., high quality and acceptable quality) will be combined prior to analysis. The purpose is to compare the data obtained from the two approaches in order to determine what quality of starting amplified DNA is necessary for successful genotyping of HLA loci.

Table 3. Amplification of HLA Exons

Singleton Amplicons:	Passed	Failed
Highest Quality	61 (85%)	11 (17%)
All Positive Amplicons	69 (96%)	3 (5%)
Multiplex Amplicons:		
Class I Exon 2	8 (100%)	------
Class I Exon 3	8 (100%)	------
Class II Exon 2	8 (100%)	------

The success in generating amplified DNA for sequencing by next generation methods is summarized in Table 2. Working with DNA from N=8 subjects nine amplicons were generated from each subject. Out of these there were 61 (85%) that upon quality control analysis resulted in a single PCR product of the correct size. In contrast, quality control analysis under the less stringent protocol indicated that 69 out of 72 (i.e., 96%) of the generated material passed. Moreover, in addition to preparation of the material by singleton PCR in which amplification targeted a single exon amplified material was also prepared using a multiplex PCR strategy (Table 3, Lower Panel). For example, exons of similar size (e.g., Class I *HLA-A*, *-B*, *-C* exon 2 and Class I *HLA-A*, *-B*, *-C* exon 3 as well as Class II *HLA-DPB*, *-DQB*, *-DRB* exon 2) were amplified in a single reaction vessel but with the appropriate combination of primers designed to co-amplify the exons. Quality control analysis of the resulting material indicated that amplified DNA of the expected size was generated (Table 3, Lower). This material is also being analyzed by next generation sequencing. An advantage of the multiplex PCR amplification approach is that, if successful, it will simplify the process of preparing material for analysis due the requirement for fewer amplification and cleanup steps.

The project timeline can be summarized. Samples from the starting cohort of N=8 subjects have been delivered to the sequencing facility. At this point we are working with the University of Pittsburgh Genomics Core Laboratory to perform the final modification steps that will occur immediately prior to next-generation sequencing. We anticipate that sequencing will be completed during the upcoming research quarter and that we will have completed our evaluation of the method's usefulness in genotyping *HLA* loci.

Statement of Plans for the Upcoming Research Period

Goal 1. Analyze results for next-generation sequencing of HLA loci. **Milestone 1A.** Analyze sequencing results for quality control, such as, quality score and read length. **Milestone 1B.** Edit sequencing raw data to enable genotyping of *HLA* loci.

Goal 2. Using the results from Goal 1 optimize protocol to enable sequencing of larger numbers of subjects. **Milestone 2A.** Estimate the number of subjects that can be multiplexed in each batch run of the next generation sequencing instrument. **Milestone 2B.** Initiate sample preparation for sequence analysis.

Literature Cited:

Cordovado SK, Zhao Y, Warram JH, Gong H, Anderson KL, Hendrix MM, Hancock LN, Cleary PA, Mueller PW. (2008) Nephropathy in type 1 diabetes is diminished in carriers of HLA-DRB1*04: the genetics of kidneys in diabetes (GoKinD) study. Diabetes 57:518-522.

Erlich H, Valdes AM, Noble J, Carlson JA, Varney M, Concannon P, Mychaleckyj JC, Todd JA, Bonella P, Fear AL, Lavant E, Louey A, Moonsamy P; Type 1 Diabetes Genetics Consortium. (2008) HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families. Diabetes 57:1084-1092.

Shendure J, Ji H. (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135-1145.

In the second quarterly scientific progress report of year 02 of our project (12/01/09 - 02/28/10), we reported on the following:

During the recently completed research quarter we have employed next generation sequencing to collect data that can be used to genotype select regions of genomic DNA. Focusing the technology on analysis of the human *HLA* region has enabled accurate genotyping of the *HLA* class I loci *HLA-A* and class II loci *HLA-DPB1*, *-DQB1*, and *-DRB1*. The research goals of the previous quarter have been met and are listed below.

Previous Quarter Research Goals

Goal 1. Analyze results from next generation sequencing of *HLA* loci.
Goal 2. Using the results from Goal 1 optimize the protocol to enable sequencing of larger numbers of subjects.

The two goals have been completed. We have obtained and analyzed sequencing data from exons 2 and 3 of *HLA-A* and exon 2 from *HLA-DPB1*, *-DQB1*, and *-DRB1* using high molecular weight DNA purified from blood and collected from n=8 subjects. Analysis of the sequences resulted in next generation sequence based genotypes that were consistent with information collected previously on these samples. Prior genotyping of the subjects was accomplished using an oligonucleotide hybridization based approach. This approach works well when moderate resolution *HLA* genotype is required, that is, when knowledge of the *HLA* allelic subgroup is sufficient for sorting subjects in broadly based categories. Human genetic studies as well as tissue/organ transplantation require high resolution *HLA* genotyping in which the exact allele is identified. Analysis of sequence-based data is required to accomplish this goal.

Detailed Description of Goals 1 and 2

Goal 1. Analyze results for next-generation sequencing of HLA loci. **Milestone 1A.** Analyze sequencing results for quality control, such as, quality score and read length.

Sequencing data were obtained using the GS-FLX next generation sequencer (454 Life Sciences, Branford, Connecticut). This instrument is capable of sequencing DNA from as many as 8,000 exons when the instrument is used in the single grid format. For our experiment we chose to work with the instrument using a subdivided grid (4 out of 8 grid wells were used) in which we repeated our sequencing assay in quadruplicate. This was done to minimize cost of the initial experiment and to collect data that can be used to judge the reproducibility, quality, and capacity of the system.

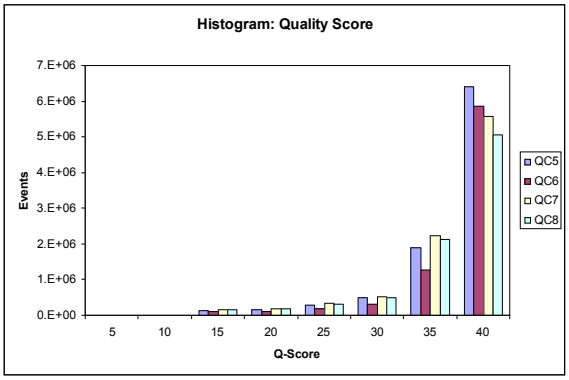
Data Quality Control and Analysis

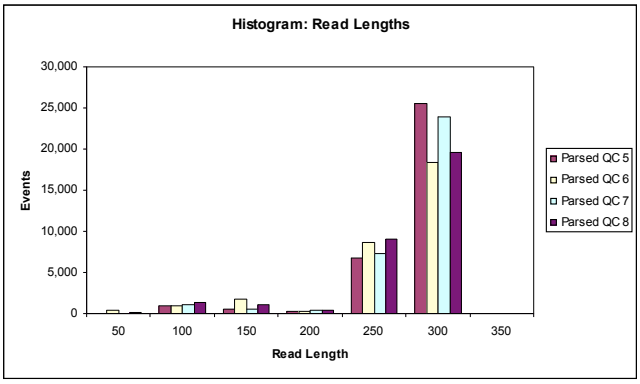
Step	Number of Sequences Remaining Grid 5	Grid 6	Grid 7	Grid 8
Processed Data	36,862 (100%)	32,891 (100%)	36,018 (100%)	34,232 (100%)
Subject/Exon ID	34,041 (92%)	30,326 (92%)	33,294 (92%)	31,473 (92%)
Read Length	32,345 (88%)	27,087 (82%)	31,222 (87%)	28,566 (83%)
Exon Start/Finish	26,959 (73%)	22,902 (70%)	25,681 (71%)	24,033 (70%)

Total Sequences Received:	140,003
Total Sequences Fail QC:	40,428 (29%)
Total Sequences Pass QC:	99,575 (71%)

The sequencing run returned data for 140,003 sequences from the combined experiment (Figure 1). The individual wells, i.e., grids 5 through 8 returned on average 35,000 sequences. This is close to the expected maximum capacity of the instrument (roughly 48,000 per grid) when the instrument is used in the 8-grid format. Moreover, we anticipated that the data would return at least 50-fold coverage for each allele sequenced. Coverage is defined as the number of times a sequence is independently obtained. It is especially useful to have coverage exceeding 25-fold when obtaining *de novo* sequencing results. This enables correction of sequencing errors due to spurious nucleotide substitutions and leads directly to the production of high quality data and consensus sequences.

The quality of the sequencing data is determined using the Q-score a measure of the confidence in calling each nucleotide base in which -log(error rate) equals the Q-score (Figure 2, left panel). The read length of each sequence was then used to identify those providing the anticipated read length greater than 200 nucleotides (Figure 2, right panel). The histogram of the Q-score (left panel) indicated that 88% of the nucleotides called exceeded a Q-score of 30, corresponding to an expected error rate of less than 1/1,000 for the majority of the data. The analysis of read length also indicated that the experiment achieved the high quality needed to call the *HLA* genotypes (Figure 2, right panel). The histogram of events versus read length indicated that 92% of the sequences exhibited reads of greater than 200 nucleotides. In these assays the protocol obtained sequence data for each target exon and on each targeted strand. The amplicons range in length from 400 to 500 base pairs. By combining sequence reads with lengths greater than 200 and Q-scores greater than 30 the analysis steps can be used to determine the sequence of the complete exon and enable accurate genotyping.





At the end of the stringent quality control process there were 99,575 or 71% of the starting sequences that passed all steps (Figure 1). Future analyses will dissect the reason(s) that 29% of the sequences failed at least one quality control criteria, however, the passed sequences represent a majority of the data and at this stage of development are deemed sufficient for completion of Milestone 1A.

Milestone 1B. Edit sequencing raw data to enable genotyping of *HLA* loci.

The sequences that passed quality control were formatted for genotyping analysis first by editing so that those generated by sequencing the reverse strand were converted to the corresponding reverse complement. This was accomplished through use of a short PERL script using the commands:

(1) `$_ = reverse($_);` Generates the reverse sequence

(2) `$_ =~ tr/ACGT/TGCA/;` Generates the complementary sequence

The data were also trimmed so that sequences corresponding to the introns were removed and exon regions were retained. This was necessary due to the use of PCR primers designed to anneal to intronic regions of the gene (for a description of the primers please see the previous quarterly report). Trimming was accomplished through use of PERL scripting to determine the index position at which each exon started (when working with the forward sequenced strand) or finished (when working with the reverse sequenced strand). The computer codes used to accomplish these steps are as below:

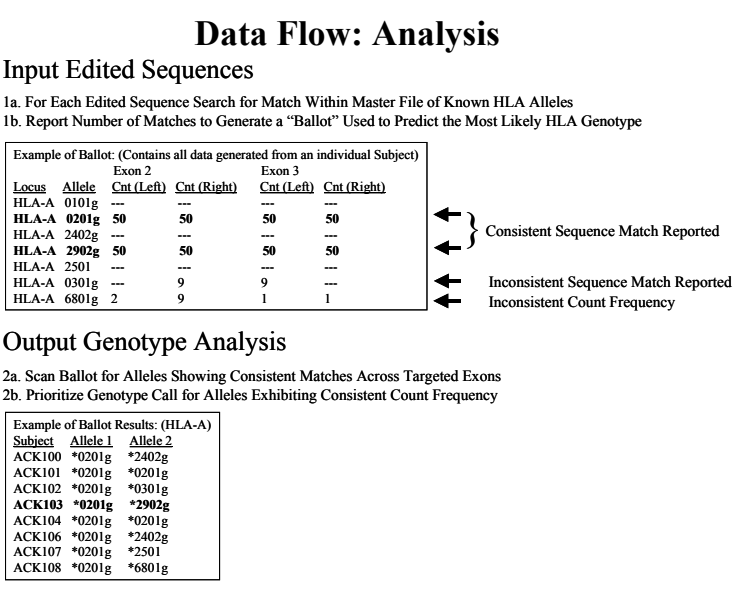
(3) $a = index($_, PATTERN); Returns the position at which exon begins or ends

(4) $_ = substring($_, $a); Returns the sequence beginning with the exon

Final editing of the sequences was performed with the goal of correcting for common insertion/deletion errors that occur when generating sequences from homopolymers greater than n=4, such as, GGGG or AAAAAA. The GS-FLX next generation sequencing instrument generates a known error in these regions in which homopolymers of 4, 5, 6 or more are poorly distinguished. To correct for this error we chose to substitute all hompolymers with 4 or more nucleotide residues to be n=3 in length. This was accomplished through use of PERL scripting and employing the command:

(5) $_ =~ s/GGGGG/GGG/g; Substitutes all homopolymer sequences for n=3

The selection of no more than 3 of the same nucleotide was determined from simulations performed *in silico* in which it was determined that the nucleotide information contained within any HLA exon could be reduced to n=3 without loss of overall sequence complexity. In other words, when sequences are uniformly substituted to n=3 for any homopolymer the HLA alleles remain unique and are able to be genotyped correctly.



The trimmed and edited sequences were used to determine the genotype of the n=8 subjects assayed in the experiment. This was preformed by generating a "ballot" of all possible allele combinations (Figure 3) and then matching the individual sequences to the dataset of all possible HLA alleles reported by the IMGT website (http://www.ebi.ac.uk/imgt/hla/). Given the total number of naturally occurring HLA alleles (determined as $(x^2 + x)/2$ where x is the number of alleles there are 694,431 possible combinations of HLA-B and 1,349,384 possible combinations for the entire set of *HLA* alleles evaluated in this study. Positive matches were chosen based upon matches to both sequencing strands (Forward and Reverse) and for both exons 2 and 3 for class I loci and exon 2 for class II loci. At the end of this process those allele genotypes in which consistent matches occurred were selected.

Cohort 1: Composite Genotypes

Subject	-A	-B	-Cw	-DPB1	-DQB1	-DRB1
ACK100	***02G1**	*270502	*010201	***020102**	***03G1**	***040101**
	***24G1**	*44G1	*050102	***040101**	***060401/0634**	***130201**
ACK101	***02G1**	*18G1	*030401	***020102**	***03G1**	***0404**
	***02G1**	*40G1	*07G1	***040101**	***030201**	***110401/02**
ACK102	***02G1**	*15G1	*030401	***030101/0502**	***03G1**	***040701/03**
	***03G1**	*18G1	***07G1**	***0402/0602**	***03G1**	***110401/02**
ACK103	***02G1**	***18G1**	***07G1**	***020102**	***03G1**	***040301/03**
	***290201**	***4404**	*160101	***030101/0502**	***030201**	***1101G1**
ACK104	***02G1**	***18G1**	***030401/03**	***040101**	***030201**	***040101**
	***02G1**	***40G1**	***07G1**	***0402/0602**	***0402**	***080101/03**
ACK106	***02G1**	*18G1	***050101/04/0503**	***020102**	***03G1**	***110401/02**
	*24G1	*44G1	***07G1**	***040101**	***050201**	***160201**
ACK107	***02G1**	*18G1	*030401	***01001**	***050101**	***010101/05**
	***250101**	*40G1	*120301	***0802/1901**	***060301**	***130101**
ACK108	*02G1	*2703	***020202**	***040101**	***030302**	***010101/05**
	***680102/6811N/6833**	*570101	*07G1	***0902/1301**	***050101**	***07010101/02**

Bold Font are consistent with expected genotype.

The summary of the genotyping results is illustrated in Figure 4. For the four grids used in the sequencing assay correct genotypes (Figure 4, bold font) were obtained for 88% of *HLA-A* (14 out of 16 alleles) and for 100% of *HLA-DPB1*, *-DQB1*, and *-DRB1* (16 out of 16). *HLA* loci *-B* and *-C* resulted in poor matches. Failing to provide convincing genotypes for any allelic combination. We anticipate that the next quarter research will focus on developing the sequencing and analysis methods and will expand upon the positive results obtained for *HLA-A* and class II loci *-DPB1*, *-DQB1*, and *-DRB1* as well as solve the difficulties observed when genotyping *-B* and *-C*.

Goal 2. Using the results from Goal 1 optimize protocol to enable sequencing of larger numbers of subjects. **Milestone 2A.** Estimate the number of subjects that can be multiplexed in each batch run of the next generation sequencing instrument.

Sequencing and Hybridization Based Genotyping

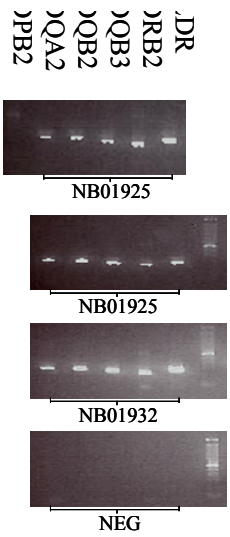
HLA-DRB1 Genotyping Results (Average Coverage >60 Sequences per Allele)

Subject	GS-FLX	Hybridization
ACK100	***040101**	***0401**/16/26/33/38
	***130201**	***1302**/39
ACK101	***0404**	***0404**/23/36/44
	***110401/02**	***1104**/36/43/44
ACK102	***040701/03**	***0407**
	***110401/02**	***1104**/43/44
ACK103	***040301/03**	***0403**/50
	***1101G1/08**	***1101**/12/15/24/27/28/29/39/49
ACK104	***040101**	***0401**/33/38
	***080101/03**	***0801**/16/26
ACK106	***110401/02**	***1104**/42/43/44
	***160201**	*1601/**02**
ACK107	***010101/05**	***0101/05**/07/08/11
	***130101**	***1301**/61
ACK108	***010101/05**	***0101/05**/07/08/11
	***07010101/02**	***0701**/03/05/07

Based upon the sequence coverage observed for *HLA-DRB1* in which each allele was observed greater than 60 times the corresponding coverage is more than adequate for *de novo* sequencing which requires greater than 25 times coverage. Figure 5 illustrates an example of the improved resolution obtained when comparing sequence based genotyping results with those obtained by the hybridization-based method. For each allele examined sequencing increased genotype resolution and resulted in a single positive exon sequence match (Figure 5, bold font).

Milestone 2B. Initiate sample preparation for sequence analysis.

Based upon the results obtained it appears likely that the approach will provide high quality data for *HLA* genotyping. We are currently in the process of preparing PCR amplicons from an independent cohort of subjects. In the new experiment we are using DNA purified from buccal swabs. These represent typical samples that we anticipate having available for future genotyping studies and provide a crucial step in the process of reducing the experimental method to practice.



Code	Oligo Name	Sequence
DPB2F	BDPBF2	GAGAGTGGCGCCTCCGCTCAT
DPB2R	BDPBR2	CCGGCCCAAAGCCCTCACTC
DQA2F	BDQAF2	GTTTCTTYCATCATTTTGTGTATTAAGGT
DQA2R	BDQAR2	CGGTAGAGTTGTAGCGTTTA
DQB2F	BDQBF2	TCCCCGCAGAGGATTTCGTGTACCA
DQB2R	BDQBR2	GACGCTCACCTCTCCGCTGCA
DQB3F	BDQBF3	TGGAGCCCACAGTGACCATCTCC
DQB3R	BDQBR3	GCTGGGGTGCTCCACGTGGCA
DRB2F	BDRBF2	CCGGATCCTTCGTGTCCCCACAGCACG
DRB2R	BDRBR2	CCGCTGCACTGTGAAGCTCTC

Figure 6 illustrates the results obtained when preparing DNA isolated from buccal swabs for amplification of the *HLA* class II loci. The panels illustrate the results of PCR amplification from n=3 subjects (upper three panels) and a no DNA negative control (bottom panel). In each example, the four *HLA* class II loci *-DPB1*, *-DQA1*, *-DQB1*, and *-DRB1* result in robust product of the expected size. This material along with similar PCR amplified material for *HLA-A*, *-B*, and *-C* are being prepared from as many as n=24 subjects. The DNA amplicons will be used during the next research quarter in order to advance the method for use with clinical samples.

Statement of Plans for the Upcoming Research Period

Goal 1. Prepare DNA collected from buccal swabs for next generation sequencing. **Milestone 1A.** Prepare amplicons for sequencing *HLA* class I loci *-A*, *-B*, *-C* and class II loci *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1*. **Milestone 1B.** Purify amplified material and pool samples to maximize capacity and throughput of the next generation sequencing approach.

Goal 2. Initiate sequencing of DNA for *HLA* class I and class II loci. **Milestone 2A.** Initiate sequencing run using the GS-FLX sequencer for massive parallel sequencing of each sample pool.

In the third quarterly scientific progress report of year 02 of our project (03/01/10 - 05/31/10), we reported on the following:

During the previous research quarter we have pursued our goal of using next generation sequencing methods to enable the development of new protocols for high throughput, accurate genotyping of genomic DNA. Our focus has been on sequencing of *HLA* exons. Thus far, we have achieved high resolution typing of class II loci *HLA-DPB*, *-DQB*, and *DRB* as well as class I loci *HLA-A*. These results were attained using high molecular weight DNA isolated from blood samples. Our work during the recently completed research quarter built upon these results by using our newly developed protocols to analyze buccal swab DNA. This DNA source originates from gently scraping of the inner mouth cheek and represents a source of DNA that while not of the highest research quality is, in fact, typical of material that we expect to encounter when genotyping human samples available from subjects enrolled in clinic-based studies.

Previous Quarter Research Goals

Goal 1. Prepare DNA collected from buccal swabs for next generation sequencing.
Goal 2. Initiate sequencing of DNA for *HLA* class I and class II loci.

Goal 1 has been partially completed while initiation of Goal 2 is awaiting the completed results from Goal 1. Briefly, our strategy for this research quarter has been to work with buccal DNA samples to determine the feasibility of PCR amplification of relevant exons from 3 class I (*HLA-A*, *-B*, *-C*) and 4 of class II (*HLA-DPB*, *-DQA*, *-DQB*, *-DRB*) loci. The results for class II are encouraging. Exon 2 has been prepared from *HLA-DPB*, *-DQA*, *-DQB*, and *-DRB* loci and exon 3 from the *-DQB* locus. At present these samples are being purified. The next step will be to check quality control by assessing the amount of each amplicon DNA followed by analysis of molecular weight by using a 0.7% agarose gel electrophoresis system.

Table 1. Primers used to generate exon specific amplicons.

Code	Primer Sequence
HLA-DPB1 exon 2	
Forward	GAGAGTGGCGCCTCCGCTCAT
Reverse	CCGGCCCAAAGCCCTCACTC
HLA-DQA1 exon 2	
Forward	GTTTCTTYCATCATTTTGTGTATTAAGGT
Reverse	CGGTAGAGTTGTAGCGTTTA
HLA-DQB1 exon 2	
Forward	TCCCCGCAGAGGATTTCGTGTACCA
Reverse	GACGCTCACCTCTCCGCTGCA
HLA-DQB1 exon 3	
Forward	TGGAGCCCACAGTGACCATCTCC
Reverse	GCTGGGGTGCTCCACGTGGCA
HLA-DRB1 exon 2	
Forward	CCGGATCCTTCGTGTCCCCACAGCACG
Reverse	CCGCTGCACTGTGAAGCTCTC

Goal 1 will be completed when the 5 exons of the class II loci are purified and pass quality control. We are currently working with samples collected from n=24 human subjects and the materials amplified from *HLA-DRB* and *-DQA* have passed all quality control steps. Goal 2 will be initiated immediately after completion of Goal 1. This will be accomplished by transferring the samples to the University of Pittsburgh Genomics Core Laboratory. That laboratory group will perform the final steps required for sequencing. Data will then be returned to us for analysis of sample genotypes.

Detailed Description of Goals 1 and 2

Goal 1. Prepare DNA collected from buccal swabs for next generation sequencing.

Milestone 1A. Prepare amplicons for sequencing *HLA* class I loci *-A*, *-B*, *-C* and class II loci *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1*. DNA samples collected from n=24 human subjects have been obtained by gentle scraping of the inner mouth cheek. DNA from these buccal samples will be purified and the amounts isolated will be quantified using the PicoGreen fluorescence assay. The samples providing greater than 10 ng per microliter genomic DNA are being used as sources of DNA for *HLA* genotyping. The first step in the sequencing process involves PCR amplification of exons from select *HLA* loci. As summarized in Table 1 oligonucleotide primers have been designed for directing specific amplification of 5 *HLA* class II exons. The PCR primer sequences are labeled as Forward or Reverse to indicate their respective orientation.

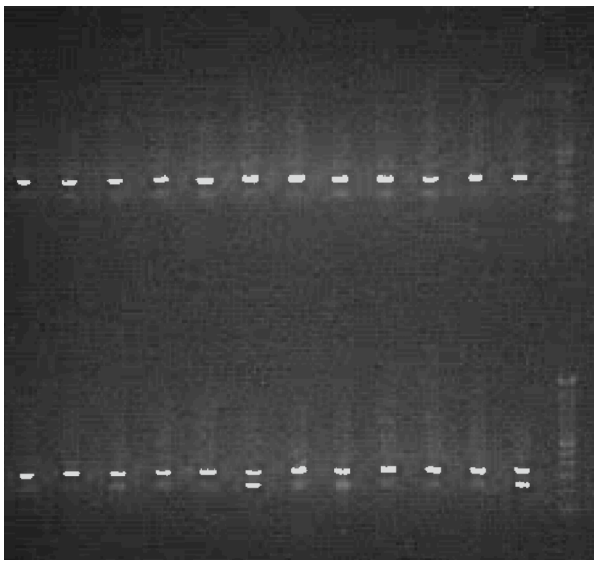
Table 2. Barcodes used to distinguish n=24 human subjects.

Subject ID	Barcode Sequence	Subject ID	Barcode Sequence
Barcode 1	ACGAGTGCGT	Barcode 13	CATAGTAGTG
Barcode 2	ACGCTCGACA	Barcode 14	CGAGAGATAC
Barcode 3	AGACGCACTC	Barcode 15	AACCAACC
Barcode 4	AGCACTGTAG	Barcode 16	AACCAAGG
Barcode 5	ATCAGACACG	Barcode 17	AACCATCG
Barcode 6	ATATCGCGAG	Barcode 18	AACCATGC
Barcode 7	CGTGTCTCTA	Barcode 19	AACCGCAT
Barcode 8	CTCGCGTGTC	Barcode 20	AACCGCTA
Barcode 9	TAGTATCAGC	Barcode 21	AACCGGAA
Barcode 10	TCTCTATGCG	Barcode 22	AACCGGTT
Barcode 11	TGATACGTCT	Barcode 23	AACCTACG
Barcode 12	TACTGAGCTA	Barcode 24	AACCTAGC

Along with the exon specific primer sequences the amplification protocol also makes use of barcoded sequences (Table 2). DNA barcodes of 8 to 10 nucleotides are used to distinguish *HLA* sequences obtained from multiple subjects. The next-generation sequencing methodology that will be used for the *HLA* genotyping project exploits a massively parallel scheme for generating sequence data. Under this experimental design multiple samples can be sequenced simultaneously. For our current experimental plan we are preparing DNA from 5 exons from 24 subjects. Assuming that most subjects in the cohort will be heterozygous for *HLA* alleles (i.e., 2 alleles per individual) the resulting 240 (=5x24x2) samples will be sequencing together. The key to unlocking the data comes during data analysis in which the careful use of exon specific primers and barcoding will enable labeling of exons and subject (see Tables 1 and 2, respectively).

Table 3. HLA Class II Amplicons.

Locus	Exon	Exon Size	Amplicon Size
HLA-DPB1	2	264	386
HLA-DQA1	2	249	373
HLA-DQB1	2	270	348
HLA-DQB1	3	282	314
HLA-DRB1	2	270	349



Milestone 1B. Purify amplified material and pool samples to maximize capacity and throughput of the next generation sequencing approach. The size of the PCR generated DNA amplicons expected are indicated in Table 3. The 5 *HLA* class II exons range in size from 249 to 282 nucleotides. Due to the addition of the barcode (Table 2) and additional sequences required during subsequent sequencing steps the final length of the amplicons are larger. For *HLA-DRB* the results of amplification are indicated in the Figure 1. The Figure shows the results when samples were treated by 0.7% agarose gel electrophoresis. By comparison of the experimentally created amplicons with a size standard of increasing 100 base pair length the expected length of the amplicons of 349 base pairs for *HLA-DRB* was confirmed. Two samples, from subjects 18 and 24, produced a second band roughly 100 base pairs smaller than expected for the correct size fragment (see subjects 18 and 24 in sample lanes 18 and 24 in Figure 1). At the moment the identity of this material is unknown. It is unlikely to be an artifact of the primer alone since the No-DNA negative controls did not exhibit a band of this size (data not shown).

Goal 2. Initiate sequencing of DNA for *HLA* class I and class II loci.

Milestone 2A. Initiate sequencing run using the GS-FLX sequencer for massive parallel sequencing of each sample pool. Sequencing of samples has not yet begun. We have taken longer than anticipated to prepare the material. The main reason for delay had to do with difficulties encountered when preparing amplicons of *HLA* class Ioci *HLA-A* and *-B*. In contrast, we have obtained excellent results when working with class II *HLA* loci (see Figure 1 as an example). These samples will be purified during the first month of the current research quarter and will provide the complete amount of materials needed to complete Goal 1 and initiate Goal 2 (sequencing). Completion of Goal 2 will be performed as the first major milestone of the current research quarter.

Statement of Plans for the Upcoming Research Period

Goal 1. Initiate sequencing of DNA for *HLA* class I and class II loci.
Milestone 1A. Initiate sequencing run using the GS-FLX sequencer for massive parallel sequencing of each sample pool.

Goal 2. Analyze sequence data obtained form class II *HLA* loci.
Milestone 2A. Perform quality control analyses of sequencing quality, length, and coverage.
Milestone 2B. Analyze sequence data that passed quality control for genotypes of *HLA* class II loci *HLA-DPB*, *-DQA*, *-DQB*, and *-DRB*.

In the fourth quarterly scientific progress report of year 02 of our project (06/01/10 - 08/31/10), we now report on our cumulative results.

The research effort of the previously completed quarter was designed to provide evidence to support the use of next-generation sequencing to genotype *HLA* class II loci *-DPB1*, *-DQA1*, *-DQB1*, and *-DRB1/3/4/5*. Goal 1 has been completed. Samples have been delivered to the University of Pittsburgh Sequencing Facility and are being analyzed on the Roche GS-FLX next-generation sequencing instrument. Goal 2, however, has not yet been completed. The purpose of the latter goal was to analyze the sequencing results for quality and then to use the best data to genotype samples. Completion of the Goal 2 will occur once the results of the sequencing run are provided.

Previous Quarter Research Goals
Goal 1. Initiate sequencing of DNA for *HLA* class I and class II loci.
Goal 2. Analyze sequence data obtained from class II *HLA* loci.

Detailed Description of Goals 1 and 2
Goal 1. Initiate sequencing of DNA for *HLA* class I and class II loci.
Milestone 1A. Initiate sequencing run using the GS-FLX sequencer for massive parallel sequencing of each sample pool. The research Goal 1 for the previously completed research quarter has been completed. Next generation sequence based genotyping of *HLA* loci is underway. Samples from n=24 human subjects have

been used to prepare DNA amplicons for exons of four *HLA* class II loci (i.e., *-DPB1*, *-DQA1*, *-DQB1*, and *-DRB1/3/4/5*). For genes *-DPB1*, *-DQA1*, and *-DRB1/3/4/5* exon 2 was prepared while for *-DQB1* exons 2 and 3 were generated as DNA template for sequencing. The primers used to PCR amplify the DNA regions have been reported in the previous quarterly report. In the recently completed research quarter we have prepared amplicons from human subjects, purified the materials, analyzed each for quality, and generated pooled samples sets of equal molar concentration.

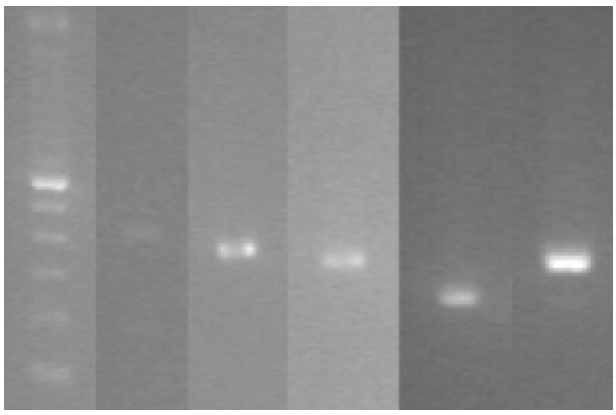


Figure 1 illustrates an example of the quality control analysis of amplicons generated from human subject NB094. Gel electrophoresis was performed using 0.7% agarose. Comparison of the DNA size standard (lane 1) with HLA class II amplicons (lanes 2 through 6) indicated that polymerase chain reaction generated amplicons were of the size expected from each of the targeted *HLA* exons. Samples were prepared from buccal DNA collected from subject NB094. Lane 1 is the 100bp marker. Lanes 2 through 6 show the amplicon generated from *HLA* loci *-DPB1* exon 2 (386bp), *-DQA1* exon 2 (373bp), *-DQB1* exon 2 (348bp), *-DQB1* exon 3 (314bp), and *-DRB1/3/4/5* exon 2 (349bp) (lanes 2 through 6, respectively). *The image shown in Figure 1 is a composite of separate gels. The buccal DNA samples were from a single individual (NB094) but different exons were analyzed on different days.*

The samples chosen for study during the recently completed research quarter were selected from previously HLA genotyped samples (Table 1). The reason for this was to incorporate a positive control into the experimental design. Analysis of the sequencing data for *HLA* genotype will be performed using computer applications developed specifically for these experiments. The prior data available on each genotype is not required for running of the application. For this reason, analysis of the sequencing data is independent of prior knowledge. Decoding of prior obtained genotypes will be done at the end of the analysis in order to test whether the next-generation sequencing method and analysis software provide similar accuracy of HLA genotyping as hybridization-based techniques used previously. For example, the genotypes for *HLA-DRB1* loci among the various human subjects used in the next-generation sequencing study are indicated in Table 1.

Table 1. Previously Determined *HLA-DRB1* Genotypes of Human Subjects

Buccal DNA Samples			Blood DNA Samples		
	HLA-DRB1 Alleles			*HLA-DRB1* Alleles	
Subject ID	Allele 1	Allele 2	Subject ID	Allele 1	Allele 2
NB094	*0401	*1201/06/10	G2932	*030101	*040101
NB099	*1302	*1501	G2936	*030101	*030101
NB106	*1502	*1502	G2963	*040101	*110401
NB108	*0701	*1501	G2983	*040101	*070101
NB110	*0402	*1501	G3017	*030101	*0404
NB113	*0404/23	*1501	G3049	*030101	*040101
NB116	*0103	*1302	G3118	*0402	*080101
NB124	*1301	*1301	G3151	030101	*040101
NB125	*0803	*1501	G3189	*030101	*040101
NB131	*0404/23	*0701	G3192	*010101/0107	*030101
NB132	*0103	*0301	G3193	*030101	*040101
NB134	*0301	*0302	G3194	*010101/0107	*030101
NB136	*0401/66	*0407	G3204	*040101	*070101
NB139	*0701	*0803	G3209	*040101	*160201
NB143	*0101	*0401	G3240	*030101	*0404
NB146	*0701	*1304	G3248	*030101	*030101
NB147	*0103	*0301	G3254	*030101	*040101

NB150	*0804	*1201/06/10	G3255	*040101	*0404
NB153	*0101	*0301	G3267	*0402	*130201
NB158	*1101	*1302	G3281	*030101	*040101
NB159	*1101	*1301	G3307	*030101	*130201
NB163	*0401	*0701	G3381	*010101/0107	*0404
NB168	*0901	*1302	G3410	*030101	*040101
NB170	*0301	*1501	G3451	*030101	*040101
NB171	*1501	*1501			
NB174	*1301	*1402			
NB178	*0301	*1503			

The experimental design of the next-generation sequencing experiment is summarized in Table 2. Samples were amplified individually and after purification their concentrations were measured using the PicoGreen fluorescence assay and the sample concentrations were adjusted to enable pooling of equal molar mixtures of as many as 24 subjects. For example in Pools 1 through 5 amplicons were pooled using 24 subjects in one exon in each pool. In Pool 6, however, all 5 exons and 24 subjects were combined to generate a pool with 120 exon derived amplicons. Pool 7 is identical to Pool 6 with the exception that 12 subjects are used, resulting in a total of 60 exon derived amplicons available for next-generation sequencing. The samples used in Pools 1 through 6 use genomic DNA collected from buccal swabs. In contrast, Pool 8 used DNA samples collected from white blood cells and was generated to genotype alleles of *HLA-DRB1/3/4/5*.

Table 2. Pooled Samples Delivered to the University of Pittsburgh Sequencing Facility.

Pool	Loci Genotyped	Exons per Subject	Number of Subjects	Total # Exons
Source of DNA is Buccal:				
1	*-DRB1/3/4/5* exon 2	1	24	24
2	*-DPB1* exon 2	1	24	24
3	*-DQA1* exon 2	1	24	24
4	*-DQB1* exon 2	1	24	24
5	*-DQB1* exon 3	1	24	24
6	Complete Class II	5	24	120
7	Complete Class II	5	12	60
Source of DNA is Blood:				
8	*-DRB1/3/4/5* exon 2	1	24	24

Goal 2. Analyze sequence data obtained from class II *HLA* loci.
Milestone 2A. Perform quality control analyses of sequencing quality, length, and coverage.
Milestone 2B. Analyze sequence data that passed quality control for genotypes of *HLA* class II loci *HLA-DPB*, *-DQA*, *-DQB*, and *-DRB*. To be completed during the beginning of the current research quarter. The DNA amplicon pools were delivered to the University of Pittsburgh Sequencing Core during the last week of July. The delay in receiving sequencing data had to do with scheduling staff scientists to begin working with our 8 pooled samples and the August vacation schedules of some staff members. The last report received from the sequencing core was that work on our project had begun and that the data will be delivered during the next few weeks.

Computer programs needed to process and perform quality control analyses on the next-generation sequencing data have been written. They have been tested and debugged during the experiments described in previous quarterly reports. Programs used previously are written in PERL and new software applications have been developed using a combination of the R and PERL programming languages. These software applications will organize the sequences by human subject and targeted *HLA* exon. The organized data will then be analyzed for read length and quality score. Sequences exhibiting read lengths exceeding 250

nucleotides and quality scores indicating sequencing accuracy of greater than 99.9% will be used for genotyping.

KEY RESEARCH ACCOMPLISHMENTS:

- Development of advanced sequencing methods for genotyping HLA loci.
- Creation of software applications for analysis of next-generation sequencing data.
- Created advanced statistical methods for analysis of SNPs and SNP-SNP pairs for association with Type 1 Diabetes
- Development of laboratory protocols for multiplex sequencing of genes associated with Type 1 Diabetes.
- Publication of 7 manuscripts.

REPORTABLE OUTCOMES:

1. Lu, L., Boehm, J., Nichol, L., Trucco, M., and Ringquist, S. Multiplex HLA typing by pyrosequencing. In Methods in Molecular Biology, vol 496: DNA and RNA Profiling in Human Blood. ed. P. Bugert. Humana Press Inc., Totowa, New Jersey (2009).

2. Kim, D.H., Ringquist, S., and Dong, H.H. Fructose - A sweet risk of fatty liver disease. In: Chocolate, Fast Foods and Sweeteners: Consumption and Health. ed. M.R. Bishop. Nova Publishers Inc., (2010).

3. Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. Screen and clean: a tool for identifying interactions in genome-wide association studies. Genetic Epidemiology 34, 275-285 (2010).

4. Kamagate, A., Kim, D.H., Zhang, T., Slusher S., Strom, S.C., Bertera, S., Ringquist, S., and Dong, H.H. FoxO1 links hepatic insulin action to endoplasmic reticulum stress. Endocrinology 151, 3521-3535 (2010).

5. Kim, D.H., Zhang, T., Ringquist, S., and Dong, H.H. Targeting FoxO1 for hypertriglyceridemia. Current Drug Targets (in press).

6. Crossett, A., Kent, B.P., Klei, L., Ringquist, S., Trucco, M., Roeder, K., and Devlin, B. Using ancestry matching to combine family-based and unrelated samples for genome-wide association studies. Annals of Statistics (in press).

CONCLUSION:

The conclusions from the final year of funding are that next-generation sequencing methods can be applied to the highly polymorphic DNA sequences found within HLA class I and class II loci. Advanced sample preparation methods improve the throughput and resolution of the reported genotypes with high accuracy and allelic resolution. Extension of the method to clinical samples show that DNA obtained from buccal swabs as well as from white blood cells can provide sufficient quality material for HLA genotyping using the next-generation sequencing methods.

The research project generated 6 publications. These are listed under the section entitled "REPORTABLE OUTCOMES".

The So What Section. What are the implications of this research? Diabetes affects 16 million Americans and 800,000 new cases annually. African, Hispanic, Native and Asian Americans are particularly susceptible to its

most severe complications. Costs associated with diabetes may be as high as $132 billion. Diabetes accounts for 42% of new cases of end-stage renal disease with over new 100,000 cases per year at an average cost of $55,000 per patient annually.

What are the military significance and public purpose of this research? As the military is a reflection of the U.S. population improved prediction of risk for developing diabetes and diabetic complications among active duty members of the military, their families, and retired military personnel will potentially allow focused preventative treatment of at risk individuals, providing significant healthcare savings and improved patient well being.

BIBLIOGRAPHY:

1. Zhang, L., Ringquist, S., Perdomo, G., Qu, S., Trucco, M., and Dong, H.H. Proteomic analysis of fructose-induced fatty liver in hamsters. Metabolism 57, 1115-1124 (2008).

2. Lu, L., Boehm, J., Nichol, L., Trucco, M., and Ringquist, S. Multiplex HLA typing by pyrosequencing. In Methods in Molecular Biology, vol 496: DNA and RNA Profiling in Human Blood. ed. P. Bugert. Humana Press Inc., Totowa, New Jersey (2009).

3. Kim, D.H., Ringquist, S., and Dong, H.H. Fructose - A sweet risk of fatty liver disease. In: Chocolate, Fast Foods and Sweeteners: Consumption and Health. ed. M.R. Bishop. Nova Publishers Inc., (2010).

4. Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. Screen and clean: a tool for identifying interactions in genome-wide association studies. Genetic Epidemiology 34, 275-285 (2010).

5. Kamagate, A., Kim, D.H., Zhang, T., Slusher S., Strom, S.C., Bertera, S., Ringquist, S., and Dong, H.H. FoxO1 links hepatic insulin action to endoplasmic reticulum stress. Endocrinology 151, 3521-3535 (2010).

6. Kim, D.H., Zhang, T., Ringquist, S., and Dong, H.H. Targeting FoxO1 for hypertriglyceridemia. Current Drug Targets (in press).

7. Crossett, A., Kent, B.P., Klei, L., Ringquist, S., Trucco, M., Roeder, K., and Devlin, B. Using ancestry matching to combine family-based and unrelated samples for genome-wide association studies. Annals of Statistics (in press).

Highlighted publications are included with this report.

Personnel receiving pay from the research effort, W81XWH-08-1-0428

2008 – 2009

Patrick Hnidka
Leah Lahoda
Robert Lakomy
Patrizia Luppi, M.D.
Steve Ringquist, Ph.D.
Eileen Roth
William Rudert, M.D., Ph.D.
Chip Scheide
Alexis Sytche
Frank Thomas
Massimo Trucco, M.D.
Catarina Wong

2009 – 2010

Patrick Hnidka
Leah Lahoda
Robert Lakomy
Patrizia Luppi, M.D.
Steve Ringquist, Ph.D.
Eileen Roth
William Rudert, M.D., Ph.D.
Chip Scheide
Alexis Sytche
Frank Thomas
Massimo Trucco, M.D.
Catarina Wong

Available online at www.sciencedirect.com
ScienceDirect
Metabolism Clinical and Experimental 57 (2008) 1115–1124
Metabolism
Clinical and Experimental
www.metabolismjournal.com

Proteomic analysis of fructose-induced fatty liver in hamsters

Lihe Zhang, German Perdomo, Dae Hyun Kim, Shen Qu, Steven Ringquist, Massimo Trucco, H. Henry Dong*

Division of Immunogenetics, Department of Pediatrics, Children's Hospital of Pittsburgh, University of Pittsburgh School of Medicine, Rangos Research Center, Pittsburgh, PA 15213, USA

Received 7 November 2007; accepted 18 March 2008

Abstract

High fructose consumption is associated with the development of fatty liver and dyslipidemia with poorly understood mechanisms. We used a matrix-assisted laser desorption/ionization–based proteomics approach to define the molecular events that link high fructose consumption to fatty liver in hamsters. Hamsters fed high-fructose diet for 8 weeks, as opposed to regular-chow–fed controls, developed hyperinsulinemia and hyperlipidemia. High-fructose–fed hamsters exhibited fat accumulation in liver. Hamsters were killed, and liver tissues were subjected to matrix-assisted laser desorption/ionization–based proteomics. This approach identified a number of proteins whose expression levels were altered by >2-fold in response to high fructose feeding. These proteins fall into 5 different categories including (1) functions in fatty acid metabolism such as fatty acid binding protein and carbamoyl-phosphate synthase; (2) proteins in cholesterol and triglyceride metabolism such as apolipoprotein A-1 and protein disulfide isomerase; (3) molecular chaperones such as GroEL, peroxiredoxin 2, and heat shock protein 70, whose functions are important for protein folding and antioxidation; (4) enzymes in fructose catabolism such as fructose-1,6-bisphosphatase and glycerol kinase; and (5) proteins with housekeeping functions such as albumin. These data provide insight into the molecular basis linking fructose-induced metabolic shift to the development of metabolic syndrome characterized by hepatic steatosis and dyslipidemia.

1. Introduction

Fructose, which occurs naturally in honey and sweet fruits, is produced in crystalline and syrup forms for commercial use. The most commonly used corn syrup contains about 55% free fructose; and its use as a sweetener in processed foods and soft drinks has greatly increased by 20% to 30% over the past 20 years, a rate of increase similar to the incidence of obesity that has risen dramatically over the same period [1]. Preclinical studies indicate that high fructose consumption is associated with the development of metabolic syndrome, as manifested by glucose intolerance, hyperinsulinemia, hypertriglyceridemia, and whole-body insulin resistance [2-6]. In addition, there are some clinical data indicating that excessive fructose consumption for a limited period predisposes healthy subjects to body weight gain with concurrent elevation in plasma triglyceride (TG) and cholesterol levels, an atherogenic lipid profile that constitutes a major risk factor for clogging the artery and causing cardiovascular disease [7-10]. Based on epidemiologic studies of obesity in relation to increased per capital consumption of high-fructose corn syrup from beverages, it is thought that excessive dietary intake of fructose is a confounding factor for the increased prevalence of overweight and morbid obesity in industrial countries [1]. There is evidence that frequent consumption of sugar-sweetened soft drinks is a potential contributing factor for childhood obesity [11-14].

Such detrimental effect of fructose on health can be ascribed to the metabolic pathway in which fructose is metabolized after its dietary intake. In this regard, fructose differs from glucose in 3 fundamental ways. First, after absorption in the gastrointestinal track, fructose fluxes via the portal circulation into the liver, where it is almost completely metabolized [15]. Unlike glucose that enters hepatocytes through glucose transporter (Glut) 2, fructose enters hepatocytes via Glut5 independently of insulin [16]. Second, glucose breakdown is negatively regulated by

* Corresponding author. Tel.: +1 412 692 6324; fax: +1 412 692 5809.
E-mail address: dongh@pitt.edu (H.H. Dong).

0026-0495/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.metabol.2008.03.017

phosphofructokinase, a hepatic enzyme that regulates glycolysis in liver, whereas fructose can evade this rate-limiting control mechanism and is metabolized into glycerol-3-phosphate and acetyl–coenzyme A. These 2 intermediate metabolites serve as substrates for glyceride synthesis, contributing to very low-density lipoprotein (VLDL)–TG production in liver [2,3]. Third, fructose, as opposed to glucose, does not directly stimulate pancreatic insulin release because of the lack of Glut5 expression in β-cells [16]. Postprandial insulin secretion is instrumental for modulating glucose metabolism in peripheral tissues and regulating energy balance via the central nervous system through both direct and indirect mechanisms to control food intake and body weight gain [17-20]. However, such an energy-balancing mechanism does not respond to dietary fructose uptake because of the inability of fructose to elicit insulin release. As a consequence, increased fructose flux into hepatocytes results in unrestrained production of intermediate metabolites, which favors energy storage by promoting de novo lipogenesis in liver.

High fructose consumption is associated with hepatic steatosis, but with poorly understood mechanisms [2-4]. To investigate the underlying mechanism of fructose-induced fatty liver, we used matrix-assisted laser desorption/ionization (MALDI)–based proteomics approach to identify candidate molecules that link high fructose consumption to the pathogenesis of hepatic steatosis. Syrian gold hamsters were fed a high-fructose diet (60% fructose, n = 6) or regular chow (n = 6) for 8 weeks. Hamsters fed on high-fructose diet, as opposed to control hamsters on regular chow, exhibited abnormal lipid profiles with increased fat deposition in liver. At the end of the 8-week treatment, hamsters were killed and liver tissues were subjected to MALDI-based proteomics. We show that high fructose feeding was associated with significant alterations in the expression of hepatic enzymes in multiple pathways. In addition to marked up-regulation of hepatic functions that promote TG synthesis and VLDL-TG production in liver, high fructose consumption resulted in perturbations in hepatic expression of antioxidant functions and molecular chaperones in protein folding. These data provide new insight into the molecular basis that links fructose-induced metabolic shift to aberrant hepatic metabolism in the pathogenesis of dyslipidemia and steatosis.

2. Materials and methods

2.1. Animal studies

Male Syrian golden hamsters (5 weeks old; body weight, 81-90 g; Charles River Laboratory, Wilmington, MA) were fed with regular rodent chow or high-fructose diet (60% fructose, DYET 161506; Dyets, Bethlehem, PA) ad libitum in sterile cages with a 12-hour light/dark cycle for 8 weeks. Blood was collected from tail vein into capillary tubes precoated with potassium-EDTA (Sarstedt, Nümbrecht, Germany) for preparation of plasma or determination of blood glucose levels using Glucometer Elite (Bayer, Mishawaka, IN). Plasma TG and cholesterol levels were determined using TG and cholesterol reagents (Thermo Electron, Melbourne, Australia). Plasma nonesterified fatty acid (NEFA) levels were determined using the Wako NEFA assay kit (Wako Chemical USA, Richmond, VA). Plasma insulin levels were determined by anti–human insulin enzyme-linked immunosorbent assay that cross-reacts with hamster insulin (ALPCO, Windham, NH). Plasma high-density lipoprotein (HDL) cholesterol levels were determined using a cardiocheck analyzer (Polymer Technology System Inc., Indianapolis, IN). Plasma non–HDL cholesterol levels were calculated as total plasma cholesterol levels minus HDL cholesterol levels. At the end of the 8-week study, hamsters were killed and liver tissues were frozen in liquid N_2. All procedures were approved by the Institutional Animal Care and Use Committee of the Children's Hospital of Pittsburgh.

2.2. Glucose tolerance test

Hamsters were fasted for 5 hours and injected intraperitoneally with 50% dextrose solution (Abbott Laboratories, Chicago, IL) at 5 g/kg body weight. Blood glucose levels were determined and plotted as a function of time. Area under the curve (AUC) of blood glucose profiles was calculated using the KaleidaGraph software (Synergy Software, Reading, PA). The AUC values are inversely correlated with the ability of hamsters to dispose intraperitoneally injected glucose.

2.3. Hepatic lipid content

Forty milligrams of liver tissue was homogenized in 800 μL of high-performance liquid chromatography–grade acetone. After incubation with agitation at room temperature overnight, aliquots (50 μL) of acetone-extract lipid suspension were used for the determination of TG concentrations using TG reagent (Thermo Electron). *Hepatic lipid content* was defined as milligram of TG per gram of liver tissue.

2.4. Liver histology

Liver tissue from euthanized animals was fixed in Histoprep tissue embedding media (Fisher Scientific, Hanover Park, IL) and snap frozen for fat staining with oil red O [21].

2.5. Liver protein extraction

Aliquots of liver tissue (40 mg) were homogenized in 800 μL of Mammalian Protein Extraction Reagent (M-PER) buffer supplemented with 8-μL protease inhibitor cocktail (Pierce, Rockford, IL). Hepatic protein extracts were obtained after centrifugation at 13 000 rpm for 10 minutes in a microfuge.

2.6. Two-dimensional fluorescence difference gel electrophoresis

Control and high-fructose–diet liver protein samples containing 300 μg protein were precipitated by 2-D Clean-Up Kit (GE Healthcare, Piscataway, NJ) and dissolved in 90 μL

lysis buffer (7 mol/L urea, 2 mol/L thiourea, 4% wt/vol CHAPS, 1% vol/vol Triton X-100, 10 mmol/L dithiothreitol). Samples were mixed with 3 μL of 100 mmol/L N-(-2-hydroxyethyl)-piperazine-N′-2-ethanesulfonic acid (HEPES) (pH 8.0). Thirty microliters of each sample was combined to create a mixed standard sample for Cy2 labeling. The standard sample was incubated with 1 nmol Cy2. The remaining aliquots of the control and high-fructose–diet samples were incubated with 1 nmol Cy3 or 1 nmol Cy5, respectively. Each labeling reaction was incubated in an ice-water bath for 20 minutes in dark. After incubation of samples, 1 μL of quenching solution (5 mol/L methylamine, pH 8.0) was added; and the mixtures were incubated on ice for an additional 30 minutes in the dark. Samples were combined and mixed with 5 μL of immobolized pH gradient (IPG) buffer and 300 μL of lysis buffer. Samples were transferred to a 1.5-mL ultracentrifuge tube and centrifuged at 100 000*g* for 20 minutes at 4°C. The supernatant was applied to an IPG strip (pH 4-7, 24 cm) and incubated for 20 hours using low voltage (30 V) in an Ettan IPGphor II IEF system (GE Healthcare). After incubation and rehydration of the IPG strip, proteins were isoelectric focused at 300 V for 30 minutes, 500 V for 30 minutes, 1000 V for 1 hour, and 8000 V for 10 hours. After isoelectric focusing, the strip was equilibrated for 15 minutes with 10 mL of 1% wt/vol dithiothreitol containing equilibration buffer (2% wt/vol sodium dodecyl sulfate [SDS], 50 mmol/L Tris-HCl pH 8.8, 6 mol/L urea, 30% vol/vol glycerol, and 0.001% bromophenol blue) and for 15 minutes with 10 mL of 2.5% wt/vol iodoacetamide containing equilibration buffer. Second-dimension SDS–polyacrylamide gel electrophoresis was performed by transferring the IPG strip to a 12.5% single-percentage gel (dimension: 1 mm, 20 cm, 26 cm) and electrophoresing with an Ettan DALT 6 electrophoresis system (GE Healthcare) for about 18 hours at 10°C.

2.7. Differential in-gel analysis

Two-dimensional (2-D) gels were scanned using a Typhoon 9400 variable mode imager (GE Healthcare). Imager settings used blue-excited fluorescence (488 nm) for Cy2, green-excited fluorescence (532 nm) for Cy3, and red-excited fluorescence (633 nm) for Cy5. Data analysis was performed using DeCyder differential analysis software, version 5.02 (GE Healthcare). Gel images were processed for spot detection and determination of the relative protein abundance based on fluorescence intensity, defined as *spot volume*. Changes in protein expression levels, expressed as spot volume ratios, were calculated after dividing the spot volume of a given protein at high-fructose conditions by its spot volume at regular chow conditions. Protein spots were selected as up-regulated or down-regulated among those exceeding a 2-fold difference in fluorescence intensity. Differentially expressed proteins were manually spot-picked from Coomassie Blue G-250 (BioRad, Hercules, CA)–stained gels, and gel plugs were transferred to 96-well collection plates.

2.8. In-gel digestion

Gel plugs were destained by washing twice with 100 μL of 50% methanol and 50 mmol/L ammonium bicarbonate at room temperature, and dehydrated with 100 μL of 100% acetonitrile for 20 minutes. Samples were transferred to 0.5-mL eppendorf tubes containing 20 μL of 100% acetonitrile and dried in a vacufuge (Eppendorf, Westbury, NY). Trypsin digestion was performed by addition of 12 μL of a 20-μg/mL trypsin solution (100 μmol/L HCl, 25 mmol/L ammonium bicarbonate, 10% acetonitrile) and incubation at 37°C overnight with gentle shaking. Supernatants were transferred to 0.5-mL eppendorf tubes; and gel plugs were extracted twice at room temperature with 50 μL of 50% acetonitrile, 1% trifluoroacetic acid (TFA) for 1 hour each extraction. Extracts were combined with the supernatant and dried in a vacufuge at room temperature. Samples were stored overnight at −20°C.

2.9. Mass spectrometry

Dried peptides from in-gel digestion were dissolved in 3 μL of 50% acetonitrile and 0.3% TFA, and mixed with 3 μL of freshly prepared matrix solution (10 mg/mL α-cyano-4-hydroxy-cinnamic acid in 50% acetonitrile, 0.3% TFA). The mixture, 0.6 μL, was spotted onto a MALDI plate (Applied Biosystems). The 4700 Proteomics Analyzer MALDI-TOF/TOF (Applied Biosystems) was used to identify proteins from the trypsin digest. Analysis of samples used reflector positive ion mode acquisition and processing method to collect peptide spectra in the mass range of 800 to 4000 d. The 10 highest-intensity peptides were selected for tandem mass spectrometry analysis using tandem mass spectrometry mode acquisition with the 1-kV positive ion and processing method. Data processing was performed with GPS Explorer Workstation (Applied Biosystems) and MASCOT database analysis of mammalian proteins.

2.10. Immunoblot assay

Aliquots (40 mg) of liver tissue were homogenized in 800 μL ice-cold M-PER solution (Pierce) supplemented with 8 μL of protease inhibitor cocktail (Pierce). Aliquots of 20 μg of protein lysates were resolved on 4% to 20% SDS–polyacrylamide gels and subjected to immunoblot assay using antibodies against chaperonin GroEL (catalog SPA-806F; Assay Designs/Stressgen Bioreagents, Ann Arbor, MI), heat shock protein 70 (Hsp70) (1:7500 dilution, catalog 3095-100; Biovision, Mountain View, CA), senescence marker protein 30 (SMP30) (1:1000 dilution, sc-25951; Santa Cruz Biotechnology, Santa Cruz, CA), protein disulfide isomerase (PDI) (1:500 dilution, 539229; Calbiochem, San Diego, CA), fatty acid binding protein (FABP) (1:15 000 dilution, NB200-434; Novus Biologicals, Littleton, CO), and apolipoprotein (apo) A-I (1:10 000 dilution, K23001R; Biodesign, Saco, ME). Proteins were detected using the chemiluminescence Western blotting reagents (Roche Diagnostics, Indianapolis, IN). The intensity of protein bands was

Table 1
Characteristics of hamsters fed on regular chow vs high-fructose diet

	Regular chow	High fructose
Body weight (g)	134 ± 4.5	142 ± 7.8
Blood glucose (mg/dL)	86 ± 12	101 ± 7
AUC (arbitrary unit)	1.0 ± 0.09	1.7 ± 0.13 *
Plasma insulin (μU/mL)	0.17 ± 0.03	0.73 ± 0.13 *
Plasma NEFA (mEq/L)	0.15 ± 0.01	0.39 ± 0.06 *
Plasma TG (mg/dL)	175 ± 20	388 ± 50 *
Plasma cholesterol (mg/dL)	149 ± 13	194 ± 16 *
Plasma HDL-C (mg/dL)	117 ± 8	151 ± 5 *
Plasma non–HDL-C (mg/dL)	31 ± 8	35 ± 5
Hepatic lipid content (mg/g liver)	6.3 ± 0.3	10.7 ± 0.8 *

Hamsters were fed regular chow or high-fructose diet for 8 weeks, followed by the determination of body weight, fasting blood glucose levels, fasting plasma levels of insulin, free fatty acid, TG, cholesterol, and HDL cholesterol (HDL-C). Non–HDL-C levels were calculated by subtracting HDL-C from total cholesterol levels in plasma. Glucose tolerance was performed after 7 weeks of fructose feeding for the determination of AUC of blood glucose profiles in response to glucose challenge. Hamsters were killed at the end of study; and liver tissues were used for the determination of *hepatic lipid content*, defined as milligram of TG per gram of wet liver tissue.

* $P < .05$ vs control by ANOVA.

quantified by densitometry using the NIH Image software (National Institutes of Health, Bethesda, MD) as described [22].

2.11. Statistics

Statistical analyses of data were performed by analysis of variance (ANOVA) using StatView software (Abacus Concepts, Berkeley, CA). The ANOVA post hoc tests were performed to study the significance between the high-fructose and regular-chow groups. Data were expressed as the mean ± SEM. P values <.05 were considered statistically significant.

3. Results and discussion

3.1. Characteristics of hamsters on regular chow vs high-fructose diet

To study the effect of high fructose consumption on glucose and lipid metabolism, we randomly assigned 5-week male hamsters into 2 groups (n = 6) to either regular chow or high-fructose diet. After 8-week feeding, we determined blood glucose and lipid parameters. As shown in Table 1, high-fructose–fed hamsters were associated with a slight body weight gain and a small increase in blood glucose levels. However, the differences in mean body weight and blood glucose levels between high-fructose and regular-chow groups did not reach a significant level, as determined by ANOVA. We also determined blood glucose profiles in response to intraperitoneal glucose challenge. Hamsters fed on high-fructose diet displayed impaired glucose tolerance, as reflected in the increased AUC values in comparison with control hamsters (Table 1). This effect mirrored the significant elevation of plasma insulin levels, which were indicative of whole-body insulin resistance in high-fructose–fed hamsters. When plasma lipid profiles were analyzed, significantly higher levels of plasma NEFA, TG, and total cholesterol were detected in high-fructose–fed hamsters. High-fructose–fed hamsters also displayed elevated HDL cholesterol levels without significant alterations in non–low-density lipoprotein cholesterol levels when compared with control hamsters. Furthermore, hamsters fed high-fructose diet exhibited significantly higher levels of hepatic lipid content. In keeping with previous observations [3,4,6,23,24], high fructose consumption is associated with lipid disorders in rodents. To corroborate these findings, hamsters were killed at the end of the 8-week study; and liver tissues were subjected to fat staining. As shown in Fig. 1, hamsters fed

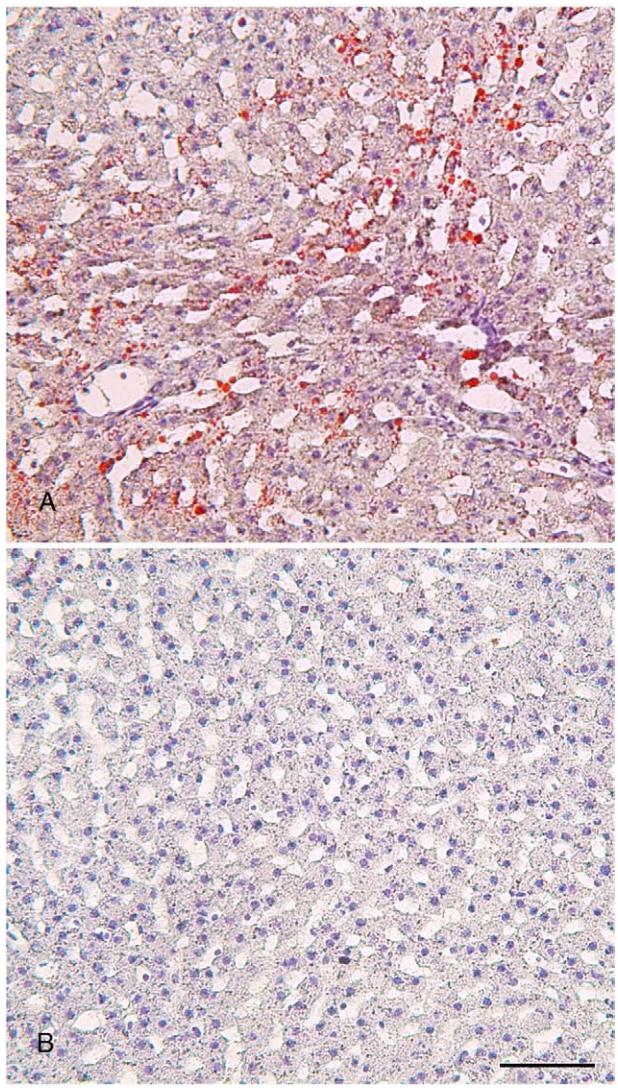


Fig. 1. Hepatic lipid content. Hamsters were killed after 8 weeks of feeding on high-fructose diet or regular chow. Liver tissues of hamsters treated with high fructose (A) and regular chow (B) were embedded with Histoprep tissue embedding media. Frozen sections (8 μm) were cut and stained with oil red O, followed by counterstaining with hematoxylin. Bar = 50 μm.

high-fructose diet were associated with increased fat deposition in liver.

3.2. Proteomic profiling of fructose-induced fatty liver

To gain insight into high-fructose–induced lipid disorder, we subjected livers of high-fructose–fed and control mice to MALDI-based proteomics because liver is the major site for fructose catabolism. As shown in Fig. 2 and Table 2, this approach identified a total of 33 protein spots whose expression levels were altered by >2-fold. These proteins fall into 5 different categories including (1) housekeeping functions such as albumin, ferritin heavy chain, and actin; (2) molecular chaperones such as GroEL and Hsp70, whose functions are important for protein folding or stress; (3) enzymes in fructose catabolism such as fructose-1,6-bisphosphatase (FBPase) and glycerol kinase (Gyk); (4) functions in lipid metabolism such as FABP and carbamoyl-phosphate synthase 1 (CPS1); and (5) proteins in cholesterol metabolism such as apo A-1. To corroborate these findings, we subjected liver tissues from control and high-fructose–fed hamsters to semiquantitative immunoblot assay. As shown in Fig. 3, this assay confirmed the results obtained from proteomics studies. Thus, in accordance with lipid disorders in high-fructose–fed hamsters, high fructose feeding resulted in significant alterations in the expression of proteins in hepatic metabolism. The physiological significance of these findings was discussed in relation to hepatic metabolism and steatosis below.

3.3. Hepatic proteins that were up-regulated in response to fructose feeding

In accordance with increased fat infiltration into liver, we detected a significant induction of FABP in high-fructose–fed hamsters (Table 2 and Fig. 3). The FABP is a cytosolic fatty acid chaperone that plays a critical role in facilitating fatty acid uptake and intracellular transport in response to dietary signals and in regulating glucose and lipid metabolism. Hepatic FABP levels are also up-regulated in response to high fat feeding or increased alcohol consumption, coinciding with the development of hepatic steatosis in mice [25,26]. In contrast, genetic ablation of hepatic FABP gene protects against high-fat–induced obesity and hepatic steatosis in mice [27-29]. These data establish FABP as an important determinant of hepatic lipid composition and turnover, suggesting that high-fructose–mediated induction of FABP production plays a causative role in increased fat deposition in livers of high-fructose–fed hamster. In support of this view, we detected a significant elevation of plasma NEFA levels in high-fructose–fed hamsters. In addition, Aoyama et al [30] showed that fructose is converted to fatty acids in liver at much greater rates than glucose. This effect, along with increased flux of fatty acids to liver, is thought to be a contributing factor for enhanced de novo lipogenesis in liver and elevated postprandial TG levels in blood in response to increased dietary fructose uptake [3,7,10,31,32].

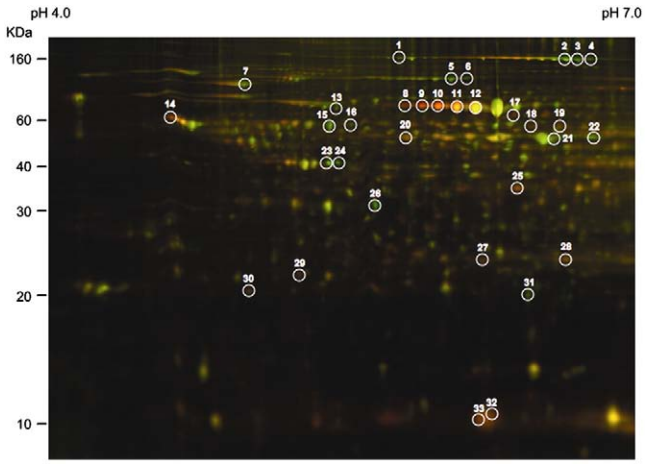


Fig. 2. Two-dimensional fluorescence difference gel electrophoretic analysis of liver proteins. In the 2-D gel image, hepatic proteins of hamsters on regular chow were shown in green color, whereas hepatic proteins of hamsters on high-fructose diet were shown in red color. Protein spots of greater than 2-fold differences between control and high fructose groups were cut out and subjected to mass spectrometry for the identification of protein ID.

Protein disulfide isomerase is an abundant multifunctional protein that resides in the lumen in the endoplasmic reticulum (ER). In response to high fructose feeding, hepatic PDI levels were markedly elevated (Table 2 and Fig. 3). The PDI functions to promote disulfide bond formation, isomerization, and reduction within the ER. In addition, PDI is associated with chaperone activities that contribute to its ability to promote proper folding of newly synthesized proteins [33-35]. In the ER, PDI forms a complex with microsomal TG transfer protein (MTP) that catalyzes the transport of TG, cholesteryl ester, and phospholipid between microsomal membranes, a rate-limiting step for hepatic VLDL assembly and secretion [36,37]. Our proteomics-based approach did not pick up MTP protein because of its relatively lower abundance in liver. However, using immunoblot assay, we and others have previously shown that hepatic MTP production was increased in hamsters in response to high fructose feeding [3,6,23,24]. This effect parallels fructose-mediated induction of PDI expression in liver, accounting in part for increased hepatic VLDL-TG production and contributing to the pathogenesis of hypertriglyceridemia in high-fructose–fed hamsters [2,3,23,38].

In response to high fructose feeding, hepatic production of apo A-1 was markedly increased (Table 2 and Fig. 3). Abundantly expressed in liver, apo A-1 is a major component of HDL and plays an important role in plasma cholesterol metabolism [39]. Apolipoprotein A-1 is necessary for the formation of nascent HDL—known as *pre-β HDL*—that acts as the acceptor of cholesterol in HDL maturation [40,41]. This effect accounts for its ability to promote reverse cholesterol transport, a dynamic process in which HDL uptakes cholesterol from peripheral tissue including macrophages for subsequent delivery to liver for

Table 2
Hepatic proteins with >2-fold alterations in response to high fructose feeding

Spot no.	Protein ID	Accession no.	Molecular mass	pI value	Score	Pattern of regulation	Volume ratio	*P* values
1	CPS1	SYRTCA	164 475	6.33	444	Down	−2.13	.05
2	CPS1	SYRTCA	164 475	6.33	573	Down	−2.51	.87
3	CPS1	SYRTCA	164 475	6.33	547	Down	−4.81	.026
4	CPS1	SYRTCA	164 475	6.33	397	Down	−4.16	.028
5	FDH	A60560	99 015	5.61	600	Down	−3.48	.00003
6	FDH	A60560	99 015	5.61	292	Down	−3.64	.0013
7	Hsp70	Q9DC41	72 378	5.01	1220	Down	−2.14	.047
8	Albumin 1	Q8C7C7	64 960	5.49	110	Up	4.66	.0023
9	Albumin 1	Q8C7C7	64 960	5.49	173	Up	5.84	.00018
10	Albumin 1	Q8C7C7	64 960	5.49	140	Up	5.39	.00079
11	Albumin 1	Q8C7C7	64 960	5.49	143	Up	4.07	.00015
12	Albumin 1	Q8C7C7	64 960	5.49	179	Up	2.79	.02
13	Annexin VI	S01786	75 838	5.34	560	Down	−2.15	.017
14	PDI	Q8R4U2	56 974	4.78	419	Up	2.77	.16
15	Chaperonin GroEL	HHMS60	60 903	5.91	1050	Down	−2.28	.25
16	BC027197 NID	AAH27197	54 014	5.69	272	Down	−2.19	.0028
17	MDH	Q921S3	63 957	6.87	132	Up	3.76	.012
18	Aldehyde dehydrogenase class 1 member B1	Q9CZS1	57 516	6.59	260	Up	2.85	.42
19	Leucine aminopeptidase	Q99P44	56 105	7.62	279	Up	2.10	.17
20	Gyk	Q8C2M1	60 522	5.47	277	Up	2.04	.0047
21	Aldehyde dehydrogenase class 2	I48966	56 501	7.53	704	Down	−2.16	.78
22	Aldehyde dehydrogenase calss 2	I48966	56 501	7.53	302	Down	−3.00	.049
23	Actin	Q61276	41 666	5.21	348	Down	−2.25	.0004
24	Actin	Q61276	41 666	5.21	231	Down	−2.71	.0005
25	FBPase	1BK4A	34 129	7.71	67	Up	2.34	.0059
26	SMP30	Q7TSW4	33 224	5.41	89	Down	−3.55	.0003
27	GST	S33860	25 953	7.71	97	Up	2.20	.1
28	GST	S33860	25 953	7.71	247	Up	2.16	.013
29	Apo A-1	Q9Z2L4	30 719	5.86	363	Up	3.35	.0013
30	PrxII	Q8K3U7	21 799	5.35	326	Up	3.07	.00005
31	Ferritin heavy chain	FRIH_CRIGR	21 341	5.73	241	Down	−3.86	.1
32	FABP	A32640	10 173	5.88	58	Up	3.01	.000001
33	FABP	A32640	10 173	5.88	66	Up	2.93	.00006

Proteomic profiling of livers of hamsters fed on high fructose (n = 6) and regular chow (n = 6) was performed. Determination of changes in protein expression levels was performed on a total of 12 2-D gels from individual hamster livers in control and fructose groups using DeCyder software version 5 and was calculated from the volume ratios of the normalized fluorescent signals. Protein spots with significant differences of greater than 2-fold ($P < .05$) between high-fructose–fed and regular-chow–fed hamsters were identified. All identified proteins match the apparent molecular mass and pI values, based on the 2-D gels.

excretion [42,43]. Reverse cholesterol transport is thought to be an important antiatherogenic mechanism for protecting against the development of atherosclerosis [42,44]. Interestingly, elevated apo A-1 production mirrors the increase in plasma HDL levels in high-fructose–induced hyperlipidemic hamsters. Likewise, Guren et al [45] showed that plasma HDL cholesterol and apo A-1 levels were elevated in obese and diabetic mice with altered lipid metabolism. Fructose-mediated induction of hepatic apo A-1 production may serve as a compensatory mechanism for increased cholesterol catabolism, as both total and HDL cholesterol levels were significantly elevated in response to high fructose feeding (Table 1).

Interestingly, we detected a marked induction of peroxiredoxin 2 (PrxII) coinciding with increased fat deposition in livers of high-fructose–fed hamsters (Table 2). This effect is accompanied by induction of the antioxidant enzyme glutathione *S*-transferase (GST) in livers in response to high fructose feeding (Table 2). Peroxiredoxin 2 is a member of the mammalian peroxiredoxin family of thiol proteins that play important roles in antioxidant defense. Expressed abundantly in liver, PrxII gene encodes a cytosolic peroxidase that functions to eliminate endogenous H_2O_2 generated from metabolism, which helps protect cells from oxidative stress and apoptosis [46,47]. Significant induction of PrxII also develops in alcohol-fed mouse livers [48]. These results raise the possibility that high fructose or alcohol consumption exerts a deleterious effect on hepatic metabolism and liver function. Fructose-mediated induction of PrxII might serve as a compensatory mechanism to alleviate the oxidant damage caused by inappropriately increased fructose catabolism in liver. In support of this notion, PrxII is abundantly expressed in liver and is markedly induced in response to ischemia/reperfusion injury during liver transplantation [49,50]. This effect has been viewed as a cytoprotective mechanism to protect liver from oxidative damage and preserve liver function posttransplantation [49,50].

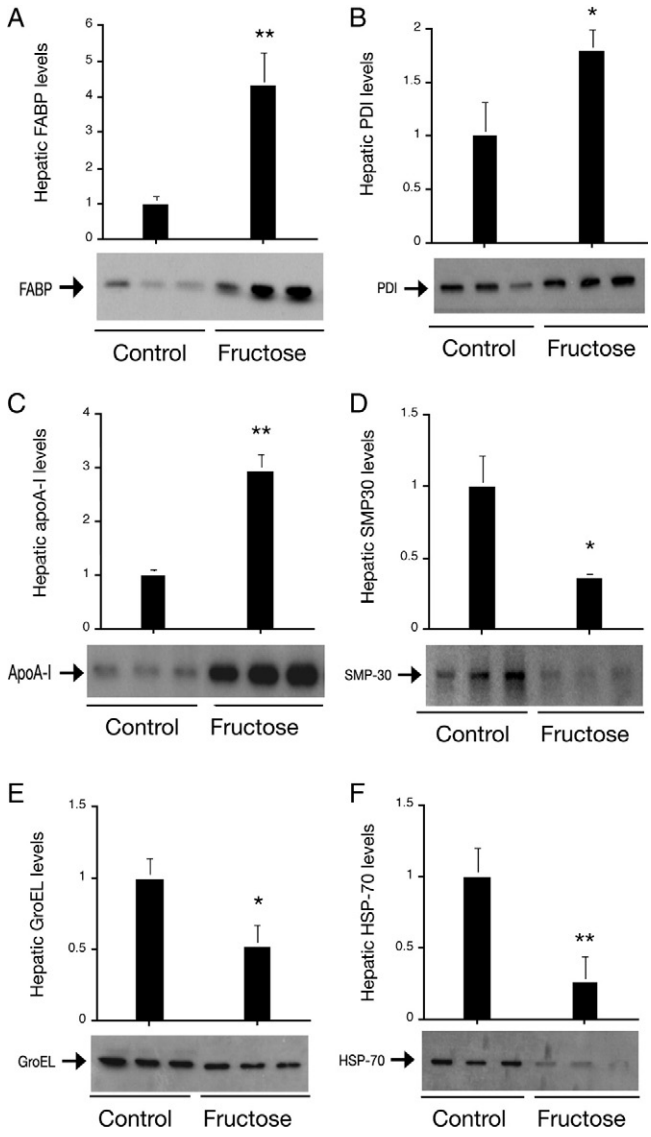


Fig. 3. Immunoblot analysis of liver proteins. Aliquots of liver tissues (40 mg) from control and high-fructose–fed hamsters were homogenized; and a fixed amount of liver proteins (20 μg) was subjected to semiquantitative immunoblot assay using antibodies against FABP (A), PDI (B), apo A-I (C), SMP30 (D), chaperonin GroEL (E), and Hsp70 (F). $^{*}P < .05$. $^{**}P < .005$ vs controls.

In addition, 2 hepatic enzymes, GyK and FBPase in glucose metabolism, were increased in response to high fructose feeding. Glycerol kinase phosphorylates glycerol to glycerol 3-phosphate, a source for dihydroxyacetone phosphate, glycerolipids, glucose, glycogen, and protein [51]. Fructose-1,6-bisphosphatase is an important gluconeogenic enzyme that catalyzes the hydrolysis of fructose 1,6-bisphosphate to fructose 6-phosphate and Pi [52]. Furthermore, hepatic levels of malate dehydrogenase (MDH) were also increased in response to high fructose feeding (Table 2). Malate dehydrogenase is an enzyme of the tricarboxylic acid cycle that converts malate and nicotinamide adenine dinucleotide (NAD) into oxaloacetate and NAD plus hydrogen (NADH), playing important roles in hepatic gluconeogenesis [53]. A potential mechanism of augmented hepatic production of GyK, FBPase, and MDH is to accommodate increased fructose catabolism and favor energy storage in high-fructose–fed hamsters (Table 2).

Leucyl aminopeptidase is another hepatic enzyme that is up-regulated in response to high fructose feeding (Table 2). Leucyl aminopeptidase plays an important role in glutathione metabolism and in the degradation of glutathione *S*-conjugates [54,55]. The physiological significance underlying fructose-mediated induction of leucyl aminopeptidase production in liver remains to be determined. In addition, we detected a significant increase in hepatic production of albumin coinciding with the increase of plasma NEFAs in high-fructose–fed hamsters (Table 2). These results are consistent with the property of serum albumin to bind and transport fatty acids in the circulation [56].

3.4. Hepatic proteins that were down-regulated in response to fructose feeding

Carbamoyl-phosphate synthase 1 is among the hepatic proteins that were down-regulated by increased fructose consumption. Carbamoyl-phosphate synthase 1 is abundantly expressed in liver and catalyzes the rate-limiting step in the urea cycle, a metabolic pathway that is primarily responsible for removing waste nitrogen from the body [57]. Hepatic deficiency of CPS1 affects the ability of liver to remove waste nitrogen, resulting in severe hyperammonemia [57]. To date, there is little information regarding the regulation of CPS1 expression in liver. Inoue et al [58] reported that genetic disruption of hepatic CCAAT/enhancer-binding protein α (C/EBPα) resulted in hepatic CPS1 deficiency, suggesting that CPS1 expression is regulated by C/EBPα in liver. We detected 2- to 4-fold reduction in hepatic CPS1 protein levels in high-fructose–fed hamsters, correlating with increased fat infiltration in liver (Table 2). These results presage an association between CPS1 deficiency and hepatic steatosis in high-fructose–fed hamsters. In support of this notion, C/EBPα null mice with inheritable CPS1 deficiency also develop age-dependent hepatic steatosis [58].

High fructose feeding also resulted in >3-fold reduction in the expression levels of 10-formyltetrahydrofolate dehydrogenase (FDH) (Table 2). 10-Formyltetrahydrofolate dehydrogenase is a high-affinity, folate-binding protein that catalyzes the $NADP^{+}$-dependent conversion of 10-formyltetrahydrofolate to CO_2 and tetrahydrofolate [59,60]. Expressed mainly in liver and brain [61-63], FDH functions to regulate the folate-mediated 1-carbon metabolism [60]. Continuous ethanol consumption in mice is associated with significantly reduced hepatic FDH activity accompanied by folate deficiency and liver weight gain [64]. The physiological significance of hepatic FDH deficiency resulting from high fructose or continuous ethanol consumption remains to be determined.

We also show that SMP30 expression in liver was significantly down-regulated by 3.5-fold in response to high fructose feeding (Table 2 and Fig. 3). The SMP30 is a 34-kd protein that is abundantly expressed in liver, lung, and kidney; and its expression levels decrease with aging [65]. The SMP30 is a lactone-hydrolyzing enzyme for biosynthesis of L-ascorbic acid, an intermediary metabolite that is involved in long-chain fatty acid metabolism in liver [66]. The SMP30 knockout mice exhibit abnormal accumulations of TGs, cholesterol, and phospholipids accompanied by an increased mortality rate [65,67]. Hepatic SMP30 levels were markedly reduced in high-fat–induced obese mice with metabolic abnormalities including hypercholesterolemia and hepatic steatosis [68]. These data together with our present studies suggest a close association that links increased fructose feeding to SMP30 deficiency, lipid disorders, and aging. Interestingly, there is evidence that continuous fructose consumption promotes the formation of advanced glycation end products and accelerates several age-related variables in male rats [69,70]. Further studies are needed to characterize the function of SMP30 in lipid metabolism and glycation for better understanding of the underlying mechanism of lipid abnormality associated with SMP30 down-regulation in liver or its potential role in aging.

Ferretins are expressed abundantly in liver and spleen, and are responsible for iron storage. Recently, Rashid et al [71] showed that ferretins interact physically with apo B in the liver. In a follow-up study, Hevi and Chuck [72] demonstrated that ferretins bind specifically to apo B and inhibit apo B secretion from cultured HepG2 cells. There is clinical evidence that a human subject with familial hypobetalipoproteinemia exhibits hepatic steatosis and liver dysfunction accompanied by marked deposition of iron in the liver [73]. Although the underlying mechanism of ferritin-mediated inhibition of hepatic apo B secretion remains to be elucidated, the available data in the literature suggest a physiological linkage between iron storage and lipid metabolism, as hepatic apo B plays a rate-limiting role in regulating TG-rich VLDL production in the liver. Consistent with this notion, we show that hepatic expression of ferritins was reduced, correlating inversely with elevated apo B and VLDL secretion in high-fructose–fed hamsters, as reported by Taghibiglou et al [74,75].

In addition to its deleterious effect on lipid metabolism, there are preclinical studies indicating that high fructose consumption is associated with oxidative stress. Rats fed a high-fructose diet exhibit increased lipid oxidation accompanied by reduced expression of antioxidant enzymes such as superoxide dismutase and glutathione peroxidase in liver and heart [76-79]. High fructose consumption also increases free radical production in rats [79,80]. Interestingly, dietary supplementation of antioxidants such as vitamin E, which mitigates oxidative stress and suppresses free radical production, ameliorates fructose-induced insulin resistance and hyperlipidemia in rats [80,81]. These data illustrate a close association between fructose-elicited oxidative stress and the development of metabolic disorders.

It is noteworthy that Morand et al [82] have used a similar proteomics approach to probe the molecular basis underlying fructose-induced hepatic insulin resistance and metabolic dyslipidemia in the hamster model. Their studies focus on the proteomic profiling of hepatic ER-associated proteins, demonstrating that high fructose consumption is associated with dysregulation of ER resident chaperones including ER60, ERp46, ERp29, PDI, and GRP94 in the liver of hamsters after 2 weeks of high fructose feeding. These ER resident proteins play important role in protein folding and lipoprotein secretion. These findings together with our present data suggest that unrestrained fructose influx into the liver result in perturbation of multiple pathways in hepatic metabolism, contributing to hepatic insulin resistance and dyslipidemia in high-fructose–fed animals.

4. Conclusion

Excessive fructose consumption is associated with dyslipidemia, culminating in markedly elevated lipid levels in plasma and increased fat deposition in liver. Our studies provide insight into the underlying mechanism of fructose-induced hepatic steatosis and diabetic dyslipidemia. We show that high fructose feeding resulted in significant alterations in multiple pathways in hepatic metabolism. These include (1) functions in fatty acid transportation, VLDL-TG assembly, and cholesterol metabolism; (2) molecular chaperones in protein folding in the ER; (3) antioxidant functions in cytoprotective mechanism; and (4) enzymes for the accommodation of fructose catabolism in response to increased fructose influx into liver. These perturbations in hepatic enzyme expressions are consistent with the idea that high fructose consumption exerts a deleterious effect on hepatic metabolism, contributing to enhanced de novo lipogenesis, augmented VLDL-TG secretion, and the development of dyslipidemia [2-4,10,83]. Although increased consumption of fructose-rich sweeteners in soft drinks is considered a contributing factor for the prevalence of obesity in industrial countries [7,8,84], our studies support the idea of limiting excessive fructose addition in beverages to counteract the epidemic of obesity and type 2 diabetes mellitus [1,84].

Acknowledgment

We thank Drs Adama Kamagate and Sandra Slusher for critical reading of this manuscript. This study was supported in part by National Health Institute grants DK066301 (HHD) and Autoimmunity Centers of Excellence U19-AI056374-01 (SR and MT), and Department of Defense ERMS 00035010 (SR and MT).

References

[1] Bray GA, Nielsen SJ, Popkin BM. Consumption of high-fructose corn syrup in beverages may play a role in the epidemic of obesity. Am J Clin Nutr 2004;79:537-43.

[2] Basciano H, Federico L, Adeli K. Fructose, insulin resistance, and metabolic dyslipidemia. Nutr Metab (Lond) 2005;2:5.

[3] Qu S, Su D, Altomonte J, et al. PPAR{alpha} mediates the hypolipidemic action of fibrates by antagonizing FoxO1. Am J Physiol Endocrinol Metab 2007;292:E421-34.

[4] Jurgens H, Haass W, Castaneda TR, et al. Consuming fructose-sweetened beverages increases body adiposity in mice. Obes Res 2005; 13:1146-56.

[5] Avramoglu RK, Qiu W, Adeli K. Mechanisms of metabolic dyslipidemia in insulin resistant states: deregulation of hepatic and intestinal lipoprotein secretion. Front Biosci 2003;8:d464-76.

[6] Taghibiglou C, Carpentier A, Van Iderstine SC, et al. Mechanisms of hepatic very low density lipoprotein overproduction in insulin resistance. Evidence for enhanced lipoprotein assembly, reduced intracellular ApoB degradation, and increased microsomal triglyceride transfer protein in a fructose-fed hamster model. J Biol Chem 2000; 275:8416-25.

[7] Teff KL, Elliott SS, Tschop M, et al. Dietary fructose reduces circulating insulin and leptin, attenuates postprandial suppression of ghrelin, and increases triglycerides in women. J Clin Endocrinol Metab 2004;89:2963-72.

[8] Elliott SS, Keim NL, Stern JS, et al. Fructose, weight gain, and the insulin resistance syndrome. Am J Clin Nutr 2002;76:911-22.

[9] Kohen-Avramoglu R, Theriault A, Adeli K. Emergence of the metabolic syndrome in childhood: an epidemiological overview and mechanistic link to dyslipidemia. Clin Biochem 2003;36: 413-20.

[10] Bantle JP, Raatz SK, Thomas W, et al. Effects of dietary fructose on plasma lipids in healthy subjects. Am J Clin Nutr 2000;72: 1128-34.

[11] Ludwig DS, Peterson KE, Gortmaker SL. Relation between consumption of sugar-sweetened drinks and childhood obesity: a prospective, observational analysis. Lancet 2001;357:505-8.

[12] James J, Kerr D. Prevention of childhood obesity by reducing soft drinks. Int J Obes 2005;29(Suppl 2):S54-7.

[13] Philippas NG, Lo CW. Childhood obesity: etiology, prevention, and treatment. Nutr Clin Care 2005;8:77-88.

[14] Gibson SA. Associations between energy density and macronutrient composition in the diets of pre-school children: sugars vs. starch. Int J Obes Relat Metab Disord 2000;24:633-8.

[15] Smith Jr LH, Ettinger RH, Seligson D. A comparison of the metabolism of fructose and glucose in hepatic disease and diabetes mellitus. J Clin Invest 1953;32:273-82.

[16] Sato Y, Ito T, Udaka N, et al. Immunohistochemical localization of facilitated-diffusion glucose transporters in rat pancreatic islets. Tissue Cell 1996;28:637-43.

[17] Schwartz MW, Porte Jr D. Diabetes, obesity, and the brain. Science 2005;307:375-9.

[18] Schwartz MW, Woods SC, Porte Jr D, et al. Central nervous system control of food intake. Nature 2000;404:661-71.

[19] Havel PJ. Control of energy homeostasis and insulin action by adipocyte hormones: leptin, acylation stimulating protein, and adiponectin. Curr Opin Lipidol 2002;13:51-9.

[20] Woods SC, Porte Jr D, Bobbioni E, et al. Insulin: its relationship to the central nervous system and to the control of food intake and body weight. Am J Clin Nutr 1985;42:1063-71.

[21] Dong H, Altomonte J, Morral N, et al. Basal insulin gene expression significantly improves conventional insulin therapy in type 1 diabetic rats. Diabetes 2002;51:130-8.

[22] Qu S, Altomonte J, Perdomo G, et al. Aberrant forkhead box O1 function is associated with impaired hepatic metabolism. Endocrinology 2006;147:5641-52.

[23] Carpentier A, Taghibiglou C, Leung N, et al. Ameliorated hepatic insulin resistance is associated with normalization of microsomal triglyceride transfer protein expression and reduction in very low density lipoprotein assembly and secretion in the fructose-fed hamster. J Biol Chem 2002;277:28795-802.

[24] Chong T, Naples M, Federico L, et al. Effect of rosuvastatin on hepatic production of apolipoprotein B–containing lipoproteins in an animal model of insulin resistance and metabolic dyslipidemia. Atherosclerosis 2006;185:21-31.

[25] Hoekstra M, Stitzinger M, van Wanrooij EJ, et al. Microarray analysis indicates an important role for FABP5 and putative novel FABPs on a Western-type diet. J Lipid Res 2006;47:2198-207.

[26] Lieber CS. Alcoholic fatty liver: its pathogenesis and mechanism of progression to inflammation and fibrosis. Alcohol 2004; 34:9-19.

[27] Newberry EP, Xie Y, Kennedy SM, et al. Protection against Western diet–induced obesity and hepatic steatosis in liver fatty acid–binding protein knockout mice. Hepatology 2006;44:1191-205.

[28] Spann NJ, Kang S, Li AC, et al. Coordinate transcriptional repression of liver fatty acid–binding protein and microsomal triglyceride transfer protein blocks hepatic very low density lipoprotein secretion without hepatosteatosis. J Biol Chem 2006; 281:33066-77.

[29] Cao H, Maeda K, Gorgun CZ, et al. Regulation of metabolic responses by adipocyte/macrophage fatty acid–binding proteins in leptin-deficient mice. Diabetes 2006;55:1915-22.

[30] Aoyama Y, Yoshida A, Ashida K. Effect of dietary fats and fatty acids on the liver lipid accumulation induced by feeding a protein-repletion diet containing fructose to protein-depleted rats. J Nutr 1974;104: 741-6.

[31] Mayes PA. Intermediary metabolism of fructose. Am J Clin Nutr 1993; 58:754S-65S.

[32] Hallfrisch J. Metabolic effects of dietary fructose. FASEB J 1990;4: 2652-60.

[33] Wetterau JR, Combs KA, McLean LR, et al. Protein disulfide isomerase appears necessary to maintain the catalytically active structure of the microsomal triglyceride transfer protein. Biochemistry 1991;30:9728-35.

[34] Wetterau JR, Aggerbeck LP, Laplaud PM, et al. Structural properties of the microsomal triglyceride-transfer protein complex. Biochemistry 1991;30:4406-12.

[35] Satoh M, Shimada A, Kashiwai A, et al. Differential cooperative enzymatic activities of protein disulfide isomerase family in protein folding. Cell Stress Chaperones 2005;10:211-20.

[36] Berriot-Varoqueaux N, Aggerbeck LP, Samson-Bouma M, et al. The role of the microsomal triglygeride transfer protein in abetalipoproteinemia. Annu Rev Nutr 2000;20:663-97.

[37] Hussain MM, Shi J, Dreizen P. Microsomal triglyceride transfer protein and its role in apoB-lipoprotein assembly. J Lipid Res 2003; 44:22-32.

[38] Guo Q, Wang PR, Milot DP, et al. Regulation of lipid metabolism and gene expression by fenofibrate in hamsters. Biochim Biophys Acta 2001;1533:220-32.

[39] Barter PJ, Rye KA. The rationale for using apoA-I as a clinical marker of cardiovascular risk. J Intern Med 2006;259: 447-54.

[40] Chau P, Nakamura Y, Fielding CJ, et al. Mechanism of prebeta-HDL formation and activation. Biochemistry 2006;45:3981-7.

[41] Rye KA, Barter PJ. Formation and metabolism of prebeta-migrating, lipid-poor apolipoprotein A-I. Arterioscler Thromb Vasc Biol 2004;24: 421-8.

[42] Lewis GF, Rader DJ. New insights into the regulation of HDL metabolism and reverse cholesterol transport. Circ Res 2005;96: 1221-32.

[43] Stein O, Ben-Naim M, Dabach Y, et al. Macrophage cholesterol efflux to free apoprotein A-I in C3H and C57BL/6 mice. Biochem Biophys Res Commun 2002;290:1376-81.

[44] Tangirala RK, Tsukamoto K, Chun SH, et al. Regression of atherosclerosis induced by liver-directed gene transfer of apolipoprotein A-I in mice. Circulation 1999;100:1816-22.
[45] Gruen ML, Plummer MR, Zhang W, et al. Persistence of high density lipoprotein particles in obese mice lacking apolipoprotein A-I. J Lipid Res 2005;46:2007-14.
[46] Low FM, Hampton MB, Peskin AV, et al. Peroxiredoxin 2 functions as a noncatalytic scavenger of low-level hydrogen peroxide in the erythrocyte. Blood 2007;109:2611-7.
[47] Yang CS, Lee DS, Song CH, et al. Roles of peroxiredoxin II in the regulation of proinflammatory responses to LPS and protection against endotoxin-induced lethal shock. J Exp Med 2007;204:583-94.
[48] Kim BJ, Hood BL, Aragon RA, et al. Increased oxidation and degradation of cytosolic proteins in alcohol-exposed mouse liver and hepatoma cells. Proteomics 2006;6:1250-60.
[49] Shau H, Merino A, Chen L, et al. Induction of peroxiredoxins in transplanted livers and demonstration of their in vitro cytoprotection activity. Antioxid Redox Signal 2000;2:347-54.
[50] Cesaratto L, Vascotto C, D'Ambrosio C, et al. Overoxidation of peroxiredoxins as an immediate and sensitive marker of oxidative stress in HepG2 cells and its application to the redox effects induced by ischemia/reperfusion in human liver. Free Radic Res 2005;39: 255-68.
[51] Herzog B, Waltner-Law M, Scott DK, et al. Characterization of the human liver fructose-1,6-bisphosphatase gene promoter. Biochem J 2000;351(Pt 2):385-92.
[52] Lamont BJ, Visinoni S, Fam BC, et al. Expression of human fructose-1,6-bisphosphatase in the liver of transgenic mice results in increased glycerol gluconeogenesis. Endocrinology 2006;147:2764-72.
[53] Kondo H, Minegishi Y, Komine Y, et al. Differential regulation of intestinal lipid metabolism-related genes in obesity-resistant A/J vs. obesity-prone C57BL/6J mice. Am J Physiol Endocrinol Metab 2006; 291:E1092-9.
[54] Josch C, Klotz LO, Sies H. Identification of cytosolic leucyl aminopeptidase (EC 3.4.11.1) as the major cysteinylglycine-hydrolysing activity in rat liver. Biol Chem 2003;384:213-8.
[55] Cappiello M, Lazzarotti A, Buono F, et al. New role for leucyl aminopeptidase in glutathione turnover. Biochem J 2004;378:35-44.
[56] Curry S, Brick P, Franks NP. Fatty acid binding to human serum albumin: new insights from crystallographic studies. Biochim Biophys Acta 1999;1441:131-40.
[57] Summar ML, Hall L, Christman B, et al. Environmentally determined genetic expression: clinical correlates with molecular variants of carbamyl phosphate synthetase I. Mol Genet Metab 2004;81(Suppl 1): S12-9.
[58] Inoue Y, Inoue J, Lambert G, et al. Disruption of hepatic C/EBPalpha results in impaired glucose tolerance and age-dependent hepatosteatosis. J Biol Chem 2004;279:44740-8.
[59] Tsybovsky Y, Donato H, Krupenko NI, et al. Crystal structures of the carboxyl terminal domain of rat 10-formyltetrahydrofolate dehydrogenase: implications for the catalytic mechanism of aldehyde dehydrogenases. Biochemistry 2007;46:2917-29.
[60] Anguera MC, Field MS, Perry C, et al. Regulation of folate-mediated one-carbon metabolism by 10-formyltetrahydrofolate dehydrogenase. J Biol Chem 2006;281:18335-42.
[61] Min H, Shane B, Stokstad EL. Identification of 10-formyltetrahydrofolate dehydrogenase-hydrolase as a major folate binding protein in liver cytosol. Biochim Biophys Acta 1988;967:348-53.
[62] Neymeyer VR, Tephly TR. Detection and quantification of 10-formyltetrahydrofolate dehydrogenase (10-FTHFDH) in rat retina, optic nerve, and brain. Life Sci 1994;54:PL395-9.
[63] Neymeyer V, Tephly TR, Miller MW. Folate and 10-formyltetrahydrofolate dehydrogenase (FDH) expression in the central nervous system of the mature rat. Brain Res 1997;766:195-204.
[64] Min H, Im ES, Seo JS, et al. Effects of chronic ethanol ingestion and folate deficiency on the activity of 10-formyltetrahydrofolate dehydrogenase in rat liver. Alcohol Clin Exp Res 2005;29:2188-93.
[65] Maruyama N, Ishigami A, Kuramoto M, et al. Senescence marker protein-30 knockout mouse as an aging model. Ann N Y Acad Sci 2004;1019:383-7.
[66] Kondo Y, Inai Y, Sato Y, et al. Senescence marker protein 30 functions as gluconolactonase in L-ascorbic acid biosynthesis, and its knockout mice areprone to scurvy. Proc Natl Acad Sci U S A 2006;103:5723-8.
[67] Ishigami A, Kondo Y, Nanba R, et al. SMP30 deficiency in mice causes an accumulation of neutral lipids and phospholipids in the liver and shortens the life span. Biochem Biophys Res Commun 2004;315: 575-80.
[68] Park JY, Seong JK, Paik YK. Proteomic analysis of diet-induced hypercholesterolemic mice. Proteomics 2004;4:514-23.
[69] Levi B, Werman MJ. Long-term fructose consumption accelerates glycation and several age-related variables in male rats. J Nutr 1998; 128:1442-9.
[70] Mikulikova K, Eckhardt A, Kunes J, et al. Advanced glycation end-product pentosidine accumulates in various tissues of rats with high fructose intake. Physiol Res 2007.
[71] Rashid KA, Hevi S, Chen Y, et al. A proteomic approach identifies proteins in hepatocytes that bind nascent apolipoprotein B. J Biol Chem 2002;277:22010-7.
[72] Hevi S, Chuck SL. Ferritins can regulate the secretion of apolipoprotein B. J Biol Chem 2003;278:31924-9.
[73] Whitfield AJ, Barrett PH, Robertson K, et al. Liver dysfunction and steatosis in familial hypobetalipoproteinemia. Clin Chem 2005;51: 266-9.
[74] Taghibiglou C, Van Iderstine SC, Kulinski A, et al. Intracellular mechanisms mediating the inhibition of apoB-containing lipoprotein synthesis and secretion in HepG2 cells by avasimibe (CI-1011), a novel acyl–coenzyme A: cholesterol acyltransferase (ACAT) inhibitor. Biochem Pharmacol 2002;63:349-60.
[75] Taghibiglou C, Rashid-Kolvear F, Van Iderstine SC, et al. Hepatic very low density lipoprotein-ApoB overproduction is associated with attenuated hepatic insulin signaling and overexpression of protein-tyrosine phosphatase 1B in a fructose-fed hamster model of insulin resistance. J Biol Chem 2002;277:793-803.
[76] Fields M, Ferretti RJ, Reiser S, et al. The severity of copper deficiency in rats is determined by the type of dietary carbohydrate. Proc Soc Exp Biol Med 1984;175:530-7.
[77] Busserolles J, Rock E, Gueux E, et al. Short-term consumption of a high-sucrose diet has a pro-oxidant effect in rats. Br J Nutr 2002;87: 337-42.
[78] Busserolles J, Zimowska W, Rock E, et al. Rats fed a high sucrose diet have altered heart antioxidant enzyme activity and gene expression. Life Sci 2002;71:1303-12.
[79] Busserolles J, Gueux E, Rock E, et al. High fructose feeding of magnesium deficient rats is associated with increased plasma triglyceride concentration and increased oxidative stress. Magnes Res 2003;16:7-12.
[80] Busserolles J, Gueux E, Rock E, et al. Substituting honey for refined carbohydrates protects rats from hypertriglyceridemic and prooxidative effects of fructose. J Nutr 2002;132:3379-82.
[81] Faure P, Rossini E, Lafond JL, et al. Vitamin E improves the free radical defense system potential and insulin sensitivity of rats fed high fructose diets. J Nutr 1997;127:103-7.
[82] Morand JP, Macri J, Adeli K. Proteomic profiling of hepatic endoplasmic reticulum–associated proteins in an animal model of insulin resistance and metabolic dyslipidemia. J Biol Chem 2005;280: 17626-33.
[83] Ostos MA, Recalde D, Baroukh N, et al. Fructose intake increases hyperlipidemia and modifies apolipoprotein expression in apolipoprotein AI-CIII-AIV transgenic mice. J Nutr 2002;132:918-23.
[84] Raben A, Vasilaras TH, Moller AC, et al. Sucrose compared with artificial sweeteners: different effects on ad libitum food intake and body weight after 10 wk of supplementation in overweight subjects. Am J Clin Nutr 2002;76:721-9.

Screen and Clean: A Tool for Identifying Interactions in Genome-Wide Association Studies

Jing Wu,[1] Bernie Devlin,[2] Steven Ringquist,[3] Massimo Trucco,[3] and Kathryn Roeder[1]*

[1]*Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania*
[2]*Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania*
[3]*Division of Immunogenetics, Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, Pennsylvania*

Epistasis could be an important source of risk for disease. How interacting loci might be discovered is an open question for genome-wide association studies (GWAS). Most researchers limit their statistical analyses to testing individual pairwise interactions (i.e., marginal tests for association). A more effective means of identifying important predictors is to fit models that include many predictors simultaneously (i.e., higher-dimensional models). We explore a procedure called screen and clean (SC) for identifying liability loci, including interactions, by using the lasso procedure, which is a model selection tool for high-dimensional regression. We approach the problem by using a varying dictionary consisting of terms to include in the model. In the first step the lasso dictionary includes only main effects. The most promising single-nucleotide polymorphisms (SNPs) are identified using a screening procedure. Next the lasso dictionary is adjusted to include these main effects and the corresponding interaction terms. Again, promising terms are identified using lasso screening. Then significant terms are identified through the cleaning process. Implementation of SC for GWAS requires algorithms to explore the complex model space induced by the many SNPs genotyped and their interactions. We propose and explore a set of algorithms and find that SC successfully controls Type I error while yielding good power to identify risk loci and their interactions. When the method is applied to data obtained from the Wellcome Trust Case Control Consortium study of Type 1 Diabetes it uncovers evidence supporting interaction within the HLA class II region as well as within Chromosome 12q24.
Genet. Epidemiol. 2010.

Key words: association test; gene-gene interaction; lasso; model selection

Additional Supporting Information may be found in the online version of this article.
Contract grant sponsor: The National Institutes of Health; Contract grant number: MH057881; Contract grant sponsor: The Department of Defense; Contract grant number: W81XWH-07-1-0619; Contract grant sponsor: The Wellcome Trust; Contract grant number: 085475.
*Correspondence to: Kathryn Roeder, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, 232 Baker Hall, Pittsburgh, PA 15213-3890. E-mail: roeder@stat.cmu.edu
Received 8 June 2009; Revised 13 August 2009; Accepted 12 September 2009
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20459

INTRODUCTION

With the advent of relatively inexpensive molecular methods for genotyping, genome-wide association studies (GWAS) have been carried out with notable success. Although the primary interest in GWAS is to identify single-nucleotide polymorphisms (SNPs) that are directly associated with a disease, there is growing evidence supporting the occurrence of epistasis and its contribution to risk for complex disease [Evans et al., 2006; Manolio and Collins, 2007]. Consequently, there is much interest in searching for interactions between two or more SNPs [Cordell, 2009]. The search for loci that interact is typically conducted in a candidate gene study or a genome-wide association study. Several strategies are available, including exhaustive searches [Marchini et al., 2005], data mining [Strobl et al., 2007], and Bayesian model selection [Zhang and Liu, 2007].

Due to the large number of potential comparisons for interactions, however, an exhaustive search involving all combinations of two or more markers across the genome is daunting [Cordell, 2009]. Exploring models fitting main effect and interactions simultaneously in the setting of GWAS are impractical or even impossible, depending on the complexity of the models evaluated. One natural way to reduce the computational load is to adopt two-stage strategies [Hoh et al., 2000; Kooperberg and LeBlanc, 2008; Marchini et al., 2005], in which a small number of promising SNPs are selected at the first stage (henceforth candidate SNPs) and the higher-order interactions are only considered among these candidate SNPs. With these strategies, the question of how to choose the promising SNPs at the first stage is crucial. Evans et al. [2006] investigate complex epistatic models and determine conditions for which it is challenging to capture the right terms to include in the second stage of a two-stage approach. It is not clear how often these conditions arise in practice because many genetic interactions demonstrate substantial marginal effects. We explore the potential of two-stage searches using a new statistical approach.

Our aim is to find a parsimonious model including SNPs, pairs of SNPs, and even higher-order interactions that best explains the phenotype. This multivariate model selection approach can improve performance over tests for

individual SNPs because it decreases the unexplained variance in the model. It has long been recognized that failing to account for these sources of heterogeneity can reduce the power to detect genetic factors in both linkage and association studies [Chatterjee et al., 2006; Hoggart et al., 2008]. In addition, a model selection approach will tend to include fewer spurious results because an SNP or multiple SNP interaction will only be included in the model if it substantially improves prediction beyond that obtained from the terms already included. A computationally efficient method for model selection is the lasso method, which is a tool for high-dimensional regression [Tibshirani, 1996].

A good model identifies a set of SNPs and interactions between SNPs (covariates) that predict the phenotype. A parsimonious model tends to err on the side of simplicity, including only a subset of predictive SNPs, while a complex model tends to include too many SNPs, some having no impact on risk. Like stepwise regression, the lasso can explore models with more covariates than observations, but the lasso is a ''less greedy'' procedure than stepwise regression in that it tends to find less complex models. As it searches the model space it can both drop and add covariates. Using a computationally efficient procedure, the algorithm returns a suite of solutions, ranging from parsimonious to complex, indexed by a complexity parameter. Thus, using the lasso, the problem of identifying genetic variants associated with the phenotype is equivalent to selecting a complexity parameter between 0 and 1. The chosen parameter identifies a single solution from among the range of solutions suggested by the algorithm. A good choice corresponds to a model that controls the Type I error rate and yet attains good power. The complexity parameter is typically chosen by statistical sampling procedures such as cross-validation. This approach yields a model with good predictive power, but the model often includes extra terms and thus it has a high Type I error rate [Devlin et al., 2003].

Due to limited research funds, or as a result of how the research unfolds, GWAS are conducted in stages. These multiple stage designs help to identify SNPs truly affecting risk by winnowing, at each stage, the list of associated SNPs. The same design feature can be used to improve the Type I error rate of the lasso.

A lasso-based procedure called screen and clean (SC) incorporates multiple stage experimental designs into the lasso procedure, attaining good power and yet controlling for spurious findings for models with only main effects [Wasserman and Roeder, 2009]. The SC procedure first screens for an inclusive model among the immense class of possibilities using the lasso, then it cleans the lasso-solution, removing terms from the model, using a traditional hypotheses testing approach. Screening is performed on stage 1 data and cleaning on stage 2 data. By exploiting the two-stage design, an optimal model can be discovered, overcoming a serious hurdle to using lasso for GWAS. SC has the important statistical feature that it finds a consistent model that controls the Type I error for main effects [Wasserman and Roeder, 2009]; if all of the SNPs are genotyped at each stage of the study, a multi-split refinement of SC is available [Meinshausen et al., 2008]. We examine the properties of this alternative.

In this paper, we follow the idea of incorporating multiple stage experimental design into the lasso procedure and expand the SC procedure to select optimal models with main and interacting effects. We focus on pairwise interactions; however, the principal of hierarchical model selection extends naturally to higher-order interactions. The extended SC procedure is able to control the Type I error and attain good power for models with interactions, just as it does for models with only main effects. SC is extended in two important ways: first, we incorporate SNP-SNP interactions; and second, we devise a computationally efficient approach to the problem that scales successfully to GWAS. The set of SNPs or SNP-SNP interactions considered at a given step of the lasso regression model is called the dictionary (D). We use a dictionary that expands and contracts at each step of development. The method is a powerful alternative to marginal methods that test each SNP or pair of SNPs individually [Kooperberg and LeBlanc, 2008; Lin, 2006]. The method builds on published multivariate regression ideas [Wu et al., 2009]. It differs in that the proposed approach controls Type I error. Competing lasso procedures do not provide valid P-values. We illustrate SC using the publicly available genome-wide data on Type 1 diabetes (T1D) data from Welcome Trust Case Control Consortium [The Welcome Trust Case Control Consortium, 2007].

METHODS

STAGES AND DICTIONARIES

For a two-stage study design N_1 subjects are genotyped at L SNPs in stage 1 and N_2 subjects are genotyped at stage 2. The purpose of the second stage is to validate the findings of stage 1. As genome-wide platforms become more cost-effective both stages are likely to yield genotypes for the whole genome. In this scenario we can use all of the data efficiently by performing multisplits of the data, repeating the SC procedure.

For simplicity we assume that the L measured SNPs are coded for the additive model ($X = 0,1$ or 2), but our results extend naturally to other genetic models.

In the interest of parsimony we use statistical interactions synonymously with SNP-SNP interactions, even though epistasis can be considerably more complex in reality [Cordell, 2002; Phillips, 2008]. Let Y be a phenotype that can be either binary or quantitative. We consider main effect models with

$$g(E[Y|X]) = \beta_0 + \sum_{j=1}^{L} \beta_j X_j, \quad (1)$$

where g is an appropriate link function. Likewise we consider interaction models with

$$g(E[Y|X]) = \beta_0 + \sum_{j=1}^{L} \beta_j X_j + \sum_{i<j;i,j=1,\ldots,L} \beta_{ij} X_i X_j. \quad (2)$$

Let $S = \{j:\beta_j \neq 0, j \in 1,\ldots,L\} \cup \{(i,j):\beta_{ij} \neq 0, (i,j) \in 1,\ldots,L\}$ be the set of terms associated with the phenotype either as main effects or interactions. We assume that the number of terms associated with the phenotype is small. Our goal is to identify these terms.

In a traditional GWAS, each SNP is tested for association and hence the dictionary (D) is the full set of SNPs that passes quality control (QC) criterion. In a multi-step statistical procedure, the dictionary can contract if we

discard terms that show little evidence of association in a previous step. Or it can also expand, if we add additional terms, such as interactions. We indicate the set of covariates in the dictionary after contraction or expansion by $\mathcal{C}(D)$ and $\mathcal{E}(D)$, respectively. In this manner the many promising avenues of a huge dictionary can be explored without directly investigating the whole space. Naturally this approach works better if the true model is hierarchical (i.e., associated interactions are accompanied by main effects). Even when the true model is not hierarchical, however, models with strong interactions often demonstrate weak main effects and hence are approximately hierarchical.

SCREEN AND CLEAN

Screen and clean illustration. A variable dictionary is critical when exploring the model given in Equation (2) because it is usually not possible to fit the full model simultaneously due to the large number of covariates. To accommodate the high-dimensional challenge, we consider a statistical procedure that employs a hierarchical search. At step 1, the dictionary consists of all SNPs $D = \{X_1, X_2, \ldots, X_L\}$ entered as main effects. Those that exceed a threshold for inclusion based on lasso screening are recorded as the candidate SNPs. At step 2, the dictionary consists of the candidate SNPs identified in step 1, plus all pairwise interactions of these terms. In summary, the dictionary contracts ($\mathcal{C}$) in step 1, and based on these results the dictionary expands ($\mathcal{E}$) to include interactions,

$$D \rightarrow \mathcal{C}(D) \rightarrow \mathcal{E}(\mathcal{C}(D)).$$

Finally, terms in the resulting dictionary are tested for association.

To illustrate the concept of a contracting and expanding dictionary in action we preview SC in two selected simulation data sets (A and B). We choose these two data sets to contrast what happens when screen happens to be too liberal (A) vs. too strict (B). The algorithm employed in this example, which is subsequently described in detail, is SC for interactions. Two data sets are drawn from the model $g(E[Y|X]) = \beta(X_5X_6 + X_{10}X_{11})$. For each model, we generate 400 individuals, each with 15 SNPs coded using an additive model $X_1, \ldots, X_{15}$.

The lasso plot (Fig. 1) displays the family of solutions provided by the lasso algorithm for data set A in the initial screen of the dictionary consisting of all 15 SNPs. Let $\tilde{\beta}_j$ and $\hat{\beta}_j$ denote the coefficient of the j'th term in the dictionary, estimated by lasso and least squares, respectively. The complexity parameter, or tuning parameter, γ is defined as $\sum_j |\tilde{\beta}_j| / \sum_j |\hat{\beta}_j|$. As the complexity parameter moves from 0 to 1, new terms are introduced to or dropped from the model. Typically the model moves from parsimonious to rich as this parameter increases. The plotted curves depict the standardized regression coefficients as a function of the complexity parameter γ. Each of them starts at 0, indicating that for $\gamma = 0$ the model includes no covariates. As γ increases, two covariates enter the model with large positive coefficients indicating that these terms have a strong positive association with the phenotype. As γ increases to 0.24, two more terms enter the model with positive coefficients. In fact, any choice of the complexity parameter between 0.24 and 0.34 will yield only the four terms actually needed to form the two interactions in the true model (5,6,10,11). At $\gamma = 0.83$ the fourth SNP is dropped from the model, but this term eventually re-enters the model with a negative coefficient. As γ increases to 1, the remaining 11 terms enter the model, but the pattern of coefficients is illogical and indicative of a model that includes several correlated terms, many of which are uninterpretable. Cross-validation yields a complexity parameter of 0.45. With this choice the model identifies 7 candidate SNPs (4 true, 3 spurious). It is typical for cross-validated screening to yield an overly rich model [Devlin et al., 2003; Wasserman and Roeder, 2009]. This is why we need the clean step of the procedure.

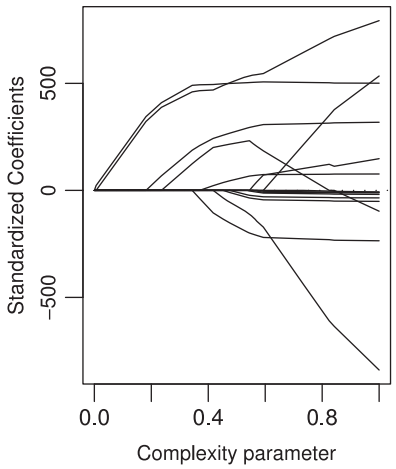


Fig. 1. Family of solutions from the lasso algorithm. As the complexity parameter increases, SNPs enter into (or drop out of) the model, one by one. Likewise, the complexity parameter determines the attenuation factor. At 0, all coefficients are 0. As the complexity parameter increases, the lasso coefficients approach the least squares solution. The traces plot each standardized coefficient as it enters the model and becomes less attenuated. Using cross-validation, a complexity parameter is selected that corresponds to a particular solution chosen from the family of solutions.

The flow chart (Fig. 2) depicts the expansion and contraction of the dictionary at each step of the SC analysis for data set A. After the initial screen, we contract the dictionary by removing all main effects not identified by the cross-validated model. We also expand the dictionary by adding all 21 pairwise interactions derived from the main effects discovered in the initial screen (Fig. 2). Applied to this dictionary, screen identifies a dictionary consisting of five interactions. Finally, using an independent sample of simulated data, the clean step of the procedure removes the three spurious interactions. The final model discovers the truth, even though the initial screen, based on cross-validation, included three spurious terms.

By chance, for set B, screen prunes the initial 15 SNP dictionary to three of the four causal SNPs, missing SNP 5 due to a lack of power. At the next step, the dictionary adds the three pairwise interaction terms corresponding to these main effects. With this dictionary of six terms, screen drops the main effects, but retains all three interactions. Finally, in the clean step, the model settles on one effect ($X_{10}X_{11}$). This is a true effect; the other true effect cannot be discovered because the model did not identify X_5 in the screen step for main effects.

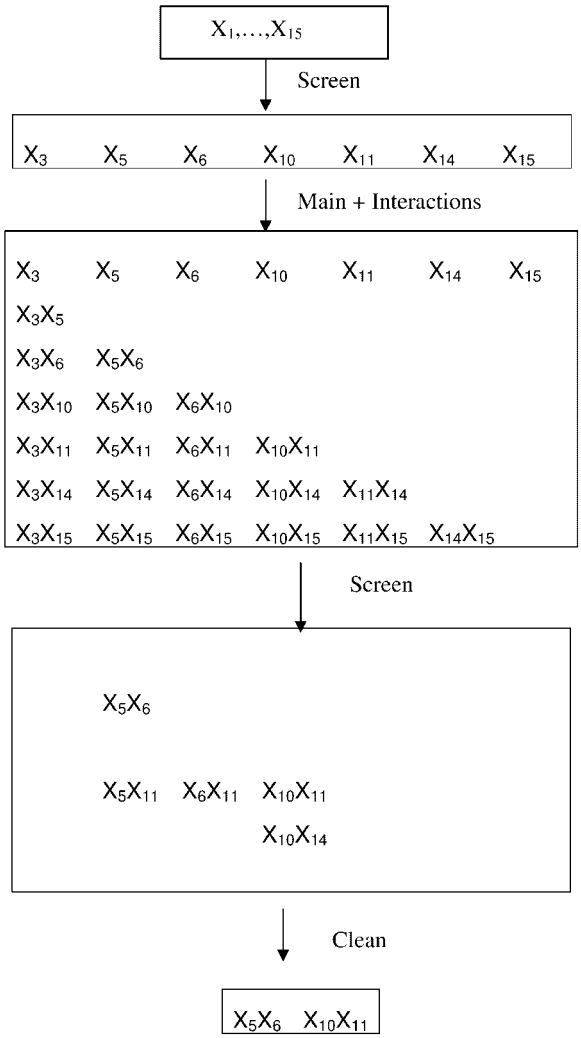


Fig. 2. Screen and clean flowchart from simulation A. Step 1: all 15 SNPs are in the model dictionary. Step 2: Screening removes all but seven terms. Step 3: Dictionary includes these seven main effects, plus pairwise interactions. Step 4: Screening removes all but five interactions. Step 5: Cleaning removes all but two interaction. This is the estimated model, which is also the true simulation model.

Screen and clean procedure. To describe the SC procedure we require notation for the number of variables under consideration. Thus we let A denote the number of variables in a set A. The SC procedure (SC_m) designed for the main effects model given in Equation (1) corresponds with a two-stage experimental design. Step 1: set the upper limit of the number of covariates to enter the screen process, L_u. This helps us to deal with the computational load; we generally set $L_u = 5{,}000$ (see more discussion in simulations). If $D > L_u$, perform marginal tests on each effect in D and select the L_u effects with the smallest p-values. Include only these terms in the revised dictionary. Step 2: using data from stage 1, the model $g(E[Y|X]) = \beta_0 + \sum_{j \in D} \beta_j X_j$ is applied to the dictionary. The lasso identifies a set of indices $\{j : \tilde{\beta}_j \neq 0\}$ for each value of the complexity parameter γ. The complexity parameter $\hat{\gamma}$ is selected by cross-validation. The resulting dictionary, $\mathcal{C}(D)$, includes all the terms for which $\tilde{\beta}_j \neq 0$ when $\gamma = \hat{\gamma}$. Step 3: using data from stage 2, find the least squares estimate $\hat{\beta}$ for the terms in $\mathcal{C}(D)$. From this analysis, obtain T_j, the traditional t-statistic obtained from the least squares analysis of the screened model, which includes $m = \mathcal{C}(D)$ terms. Clean the model of superfluous terms by selecting $\{j \in \mathcal{C}(D) : |T_j| > c\}$, in which $c = Z_{\alpha/(2m)}$.

To discover interactions as in equation (2), we extend the SC_m algorithm to handle dictionaries with interactions. We call the procedure SC_i. Repeat Steps 1 and 2 as for SC_m. Step 3: obtain $\mathcal{E}(\mathcal{C}(D))$. If $\mathcal{E}(\mathcal{C}(D)) > L_u$, perform marginal tests on each term in $\mathcal{E}(\mathcal{C}(D))$ and update $\mathcal{E}(\mathcal{C}(D))$ to include only the L_u terms with the smallest p-values. Again using data from stage 1, fit a model including all of the main effects and interactions delineated by these L_u best terms. For each $\gamma \in (0, 1)$ we obtain a contracted dictionary including $\{j : \tilde{\beta}_j \neq 0\} \cup \{(i, j) : \tilde{\beta}_{ij} \neq 0\}$. Select $\hat{\gamma}$ by leave-one-out cross-validation, and use $\hat{\gamma}$ to define the dictionary, $S = \mathcal{C}(\mathcal{E}(\mathcal{C}(D)))$, to be used for the final step; let $m^* = S$. Step 4: using the second stage data, clean the model as follows. Find the least squares estimate $\hat{\beta}$ for the model defined by S. The chosen model is $\{j : |T_j| > c\} \cup \{(i, j) : |T_{ij}| > c\}$, where $c = Z_{\alpha/(2m^*)}$, and T_j and T_{ij} are the t-statistics for main effects and interactions, respectively. The resulting procedure is designed to control Type I error at level α.

Genome-wide association. Multivariate methods such as SC do not scale directly to the immense computational burden imposed by a GWAS (results not shown). SC is computationally challenged by large numbers of covariates (p) and large numbers of subjects (n). With $n = 400$ and $p = 1{,}000$, the procedure takes less than 1 min to perform, but as the number of covariates increases to 5,000 and n increases to 1,000, the procedure requires about an hour. Approached directly, the computational challenge for hundreds of thousands of SNPs is prohibitive; however, this does not prevent us from employing SC in a GWAS. When the dimension of the problem is large we adjust the algorithm to obtain the results in a reasonable time. The adjustments include pre-selection of SNPs to those with promising marginal signals, and reducing the effort involved to perform cross-validation. These adjustments can be combined together or used individually.

Prescreening can be used to limit the number of SNPs in the dictionary. Based on a marginal test of association, most of the SNPs can be eliminated from consideration. We suggest prescreening the dictionary to include only those SNPs with a marginal p-value less than p_0. Because SC is based on a two-stage process, prescreening has no impact on the Type I error of the procedure. In addition, the number of SNPs entered in SC can be reduced by restricting the analysis to tag SNPs [de Bakker et al., 2005; Rinaldo et al., 2005].

The computational effort increases quadratically as a function of n and p. Consequently we view $p \approx 5,000$ as a practical upper limit on the number of covariates for $n \approx 2000$. This is due to two computational features in SC. Like a stepwise procedure, the lasso searches the covariate sets by adding and dropping covariates sequentially. The default maximum number of steps taken by the lasso algorithm increases with the number of samples and the number of variables in the model. This default value can be lowered to obtain an approximate solution; however, the algorithm might then fail to discover some subtle signals in the data. Second, the computational cost of the

cross-validation increases linearly in the number of samples when using leave-one-out cross-validation. For large n we suggest k-fold cross-validation, which leaves n/k observations out in each step of the algorithm instead of leaving one out [Hastie et al., 2001]. We obtain good results using k of 30–40.

Here we summarize the SC_i algorithm, with adaptations that facilitate analysis of GWAS.

SC_i algorithm.

1. Create a dictionary D including all SNPs with minor allele frequency (MAF) >0.01. To ensure that $D<L_u$, restrict this set by including only
   (a) those SNPs with marginal p-values $<p_{0m}$;
   (b) tag SNPs.
2. Using stage 1 data, screen D for main effects to obtain $\mathcal{C}(D)$. In cross-validation, restrict the class of models to those with R_1 or fewer terms.
3. Obtain $\mathcal{E}(\mathcal{C}(D))$ by including pairwise interactions. Optionally restrict this set by
   (a) including only those interactions with marginal p-value $<p_{0i}$.
4. Screen $\mathcal{E}(\mathcal{C}(D))$ to obtain $S=\mathcal{C}(\mathcal{E}(\mathcal{C}(D)))$. Again, in cross-validation, restrict the class of models to those with R_2 or fewer terms.
5. Using stage 2 data, clean S.
6. The final model includes those terms with cleaned p-values $<\alpha$. These p-values have been corrected for multiple testing.

Multi-split SC. When genotypes for the full panel of SNPs are available for every individual in the data there is no obvious split of the data into one set for screening and another set for cleaning. In this scenario, the SC analysis results vary depending on how this single-split is chosen. For this scenario, Meinshausen et al. [2008] extended the single-split SC procedure to a multi-split procedure. The analysis involves randomly splitting the data repeatedly, running SC for each split to obtain a set of p-values for each covariate, and then obtaining a single composite p-value from the sample of p-values.

For $b=1,\ldots,B$,

1. randomly split the data into two portions: $D_1^{(b)}$ for screen and $D_2^{(b)}$ for clean;
2. using $D_1^{(b)}$, screen to find the variables, $S^{(b)}$, with $\tilde{\beta}\neq 0$;
3. clean using $D_2^{(b)}$;
   (a) based on the results of clean, compute the p-values $\tilde{P}_j^{(b)}$ for variables in $S^{(b)}$;
   (b) set $\tilde{P}_j^{(b)}=1$ for variables not in $S^{(b)}$;
4. obtain a p-value that is corrected for multiple testing

$$P_j^{(b)}=\min(\tilde{P}_j^{(b)}\,|S^{(b)}|,1).$$

Thus far, the algorithm is the usual SC procedure applied repeatedly over B splits of the data. Typically a variant associated to the phenotype will produce a distribution of $P_j^{(b)}$, $b=1,\ldots,B$ including several small p-values and several 1's. Thus we cannot obtain a single p-value for each variant by taking the mean of the $P_j^{(b)}$'s. The alternative is to examine the distribution of p-values, from which a summary p-value can be obtained. Meinshausen et al. [2008] recommend the following algorithm that provides a conservative overall p-value:

1. obtain the empirical quantile function q_δ for δ in the interval [0.05, 1];
2. find δ^* to minimize the function q_δ/δ;
3. set $P_j=\min(4\times q_{\delta^*}/\delta^*,1)$.

The multiplication by 4 accounts for selecting the quantile that yields the smallest p-value [Meinshausen et al., 2008].

Other features. To control for confounding effects of population structure we suggest including eigenvectors estimated using either principal component analysis [Price et al., 2006] or spectral analysis [Lee et al., 2010]. For case-control data there is also the option of matching cases and controls by estimated ancestry and using the conditional logit model on the matched strata [Luca et al., 2008].

The lasso is designed for linear regression and quantitative traits. For dichotomous traits, logistic regression replaces linear regression naturally at a number of steps in the algorithm. This works conveniently for univariate p-values and cleaning of the data. When screening a large model space the computational challenge is greater for logistic regression. Wu et al. [2009] describe an approach that they call lasso penalized logistic regression. Following the classification literature [Hastie et al., 2001], we have found that linear regression provides a practical alternative to logistic regression even when the response variable is binary.

RESULTS

SIMULATION RESULTS WITH A MODERATE NUMBER OF SNPS

We generate 400 individuals, each with 1,000 SNPs with genotypes encoded by having $X=0,1$, or 2 minor alleles. We use half of the samples to screen (stage 1) and the remaining half to clean (stage 2). The SNPs are block-wise dependent with 200 blocks of size 5. Linkage disequilibrium within blocks is set low (Pearson's correlation coefficient [Devlin and Risch, 1995] $\rho=0.25$) and high ($\rho=0.75$). We generate a quantitative phenotype Y according to four models, with random error $\varepsilon\sim N(0,1)$. Models M1, M2, and M3 each contain multiplicative interaction terms with varying numbers of SNP-SNP pairs involved in the interaction. For ease of exposition, the coefficient β is constant for each term in all models, but for model M3 the strength of the association decreases by a multiplier for each successive SNP pair.

M0. $Y=\beta(X_5+X_6)+\varepsilon$
M1. $Y=\beta X_5X_6+\varepsilon$
M2. $Y=\beta(X_5X_6+X_{10}X_{11})+\varepsilon$
M3. $Y=\beta(X_5X_6+0.8X_{10}X_{11}+0.6X_{15}X_{16}+0.4X_{20}X_{21}+0.2X_{25}X_{26})+\varepsilon$.

For computational efficiency we performed our simulations using a relatively small sample size (400) and a large genetic signal (β ranged from 0.25 to 2.0). In regard to their power, these choices are statistically equivalent to a more realistic scenario with sample size 1,500 and genetic heritability attributable to each interaction ranging from

0.1 to 7%. We performed 1,000 simulations for each combination of model, β, and ρ.

Define power as the fraction of discoveries of interactions over the total number of interactions in the model; the false discovery rate (FDR) as the fraction of false discoveries of interactions among the total number of discoveries of interactions; and the Type I error rate as the fraction of the simulations with at least one false discovery over the total number of simulations.

We evaluate the Type I error of SC_m and SC_i, using data simulated based on model M0 (Table I). SC_m successfully controls the Type I error for each condition explored. When the marginal effects become more substantial, the Type I error of SC_i increases slightly over the nominal level.

For most configurations of parameters in models M1–M3, SC_i controls Type I error well (Fig. 3; dashed lines). The procedure has low power for small β, but power increases rapidly as the signal grows. Comparing panels moving left to right it is apparent that more complex models have lower power than simpler ones. Poor performance for SC_i occurs when both the block correlation and the model complexity are high (bottom-right panel). This suggests that better performance might be obtained if the analysis were performed using only tag SNPs.

Next, for the same conditions, we compare the performance of the method using a single-split vs. the multi-split procedure (Fig. 3; dashed vs. solid lines). For this scenario, we performed 200 simulations with SC_i repeated five times. For all three models the Type I error improved

TABLE I. Type I error in model M0 for screen and clean for main effects only (SC_m), screen and clean for main effects and interactions (SC_i)

	β = 0.5		β = 1		β = 2	
Type I	SC_m	SC_i	SC_m	SC_i	SC_m	SC_i
ρ = 0	0.06	0.06	0.06	0.07	0.05	0.11
ρ = 0.25	0.05	0.04	0.04	0.09	0.05	0.10
ρ = 0.75	0.06	0.04	0.07	0.05	0.05	0.09

Levels of disequilibrium (ρ) and strength of association (β) vary as indicated.

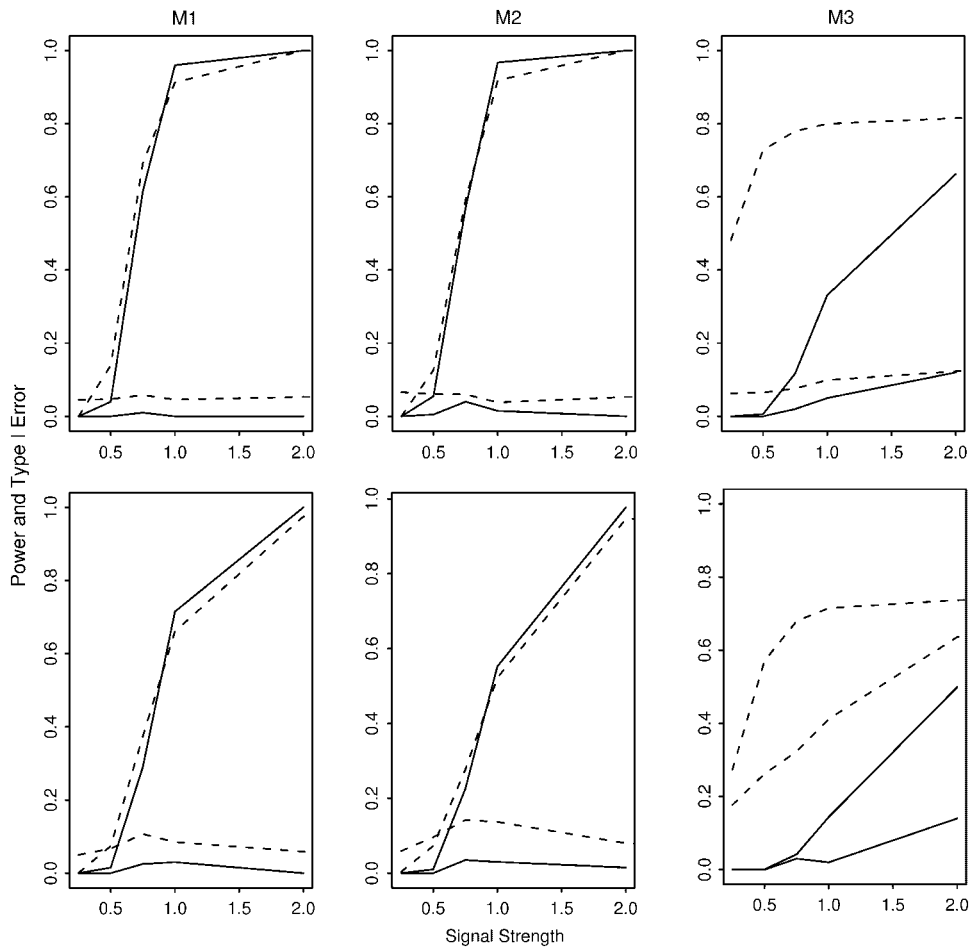


Fig. 3. Power and Type I error rates for the (one-split) SC method and the multi-split SC method. Plotted against the strength of the signal (β) are power (top two lines) and Type I error (bottom two lines) for one-split SC_i (dashed lines) and the multi-split SC_i (solid line). Top (bottom) row is low (high) correlation within blocks. Columns correspond to models of increasing complexity (M1, M2, and M3).

substantially with the multi-split procedure. For model M3 controlling the Type I error, came at the cost of a substantial loss in power, especially when the SNPs had a higher correlation. Moreover, the Type I error is still slightly inflated when the correlation is high.

To combat this problem we repeated the experiment on model M3 using tag SNPs ($\rho < 0.1$). The power is essentially unchanged from when all SNPs were included, but Type I error is successfully controlled (results not shown).

SIMULATIONS WITH LARGE NUMBERS OF SNPS

To demonstrate the application of SC to GWAS, we simulated data sets with 1,500 samples and 100,000 SNPs. The SNPs are generated to simulate tag SNPs that possess LD structure similar to a Markov chain: nearest neighbor SNPs have correlation $\rho = 0.3$. We set MAF at 0.3; in practice SNPs with a smaller MAF will require a bigger signal to yield the same power. We use two-thirds of the samples to screen (stage 1) and the remaining one-third to clean (stage 2). (This is just one option for splitting the data. Using a greater fraction of the data for cleaning might be advantageous.)

To assess the power of SC in GWAS we simulated 100 data sets for two more complex models. For model M4, 100 causal SNPs fall into sets of 10 SNPs of equal signal strength, with 10 levels ranging from low to high signal. For model M5, 25 pairs of SNPs are grouped into sets of 5, with strength of interaction signal set at five levels ranging from low to high. We use simulations from model M4 to evaluate tests for main effects and model M5 for interaction effects.

For model M4, let $X\{S_j\} = \{X_{j_1}, \ldots, X_{j_{10}}\}$, $j = 1, \ldots, 10$, represent 10 sets of 10 randomly selected SNPs for each simulation. For each j, the effect size is $j \times \beta$, so that, in total, 100 SNPs affect Y, including 10 SNPs at each level, with $\beta = 0.3$, i.e.,

$$M4 : Y = \beta \sum_{j=1}^{10} jX\{S\} + \varepsilon.$$

Assignment of the 100 causal SNPs vary by simulation.

For model M5, let $X\{T_j\} = \{(X_{j_1}X_{j_1}), \ldots, (X_{j_5}X_{j_5})\}$, $j = 1, \ldots, 5$, represent five sets of five randomly selected pairs of SNPs for each simulation. For each j, the effect size is $j \times \beta$, $j = 1, \ldots, 5$, so that, in total, 25 SNPs affect Y, including five pairs of SNPs for each level, with $\beta = 0.9$, i.e.,

$$M5 : Y = \beta \sum_{j=1}^{5} jX\{T_j\} + \varepsilon.$$

Again the 25 pairs of causal SNPs vary by simulation.

For each simulated data set, using stage 1 data, we first prescreen the SNPs and include only those SNPs with a marginal p-value less than 0.05, effectively reducing the size of the dictionary to approximately 5,000. In screen, we use the default parameters in lasso and leave-one-out cross-validation and otherwise follow the described procedure for SC in GWAS.

For main effects SC_m achieves reasonable control of both Type I error (0.077) and the FDR (0.0016). For interactions, SC_i has a higher than desired Type I error (0.13), but these errors are fairly uncommon compared to true discoveries,

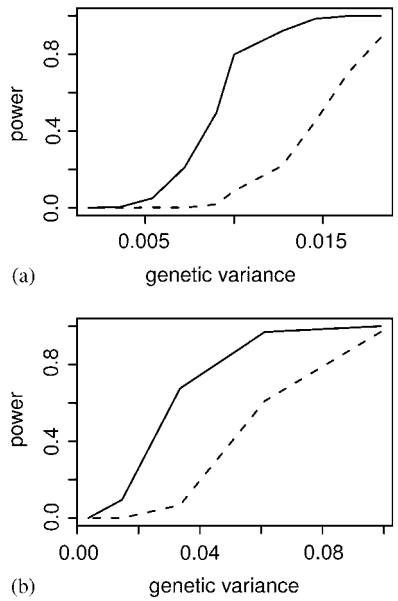


Fig. 4. Power for SC_m (a) and SC_i (b) methods. Genetic variation is the average fraction of the total genetic variance attributable to a SNP (a) or an SNP-SNP interaction (b).

as evidenced by the well controlled FDR (0.014). To assess power, we used SC_m on Model M4 and SC_i on Model M5 (Fig. 4). Notice that the power to detect main effects is much greater than the power to detect interactions.

To assess the advantages of multivariate model selection we compared SC with methods designed for marginal testing of main effects [Lin, 2006] and interactions [Kooperberg and LeBlanc, 2008]. Both marginal methods use the first stage data to screen for important main effects. The second stage data is then combined with the first stage data for a test of each effect selected in the screening process. In a test for main effects, Lin's method had false-positive rates comparable to SC_m (Type I error = 0.088, FDR = 0.003), but substantially lower power (Fig. 4, top panel). Kooperberg and LeBlanc's (KL) method tests for interactions formed from all pairwise combinations of screened main effects. This method showed false-positive rates equivalent to SC_i (Type I = 0.18, FDR = 0.022), but considerably lower power (Fig. 4, bottom panel).

We also explored some of the simpler non-additive models investigated by Evans et al. [2006]. We chose these models because they are derived from pairwise combinations of recessive and dominant single SNP models. The recessive-recessive (RR) only has an effect if all four minor alleles are present; the recessive-dominant (RD) has an effect if SNP one has both minor alleles and SNP two has at least one minor allele; the dominant-dominant (DD) has an effect if both SNPs have at least one minor allele; and the dominant-recessive-dominant (DRD), has the effect if at least three minor alleles are present. A two-step analysis is likely to fail if the main effects explain an insufficient fraction of the variance contained in the interaction. For this set of models, the fraction of the variance attributable to main effects and epistatic effects for each varies (Table II). For comparison we include a model with the core element we used for most of our simulations $Y = \beta X_1 X_2 + \varepsilon$; we label this model M, because it is based on a multiplicative

TABLE II. The fraction of the genetic variance attributable to each main effect and the epistatic effect in five genetic models

Model	RR	RD	DRD	DD	M
Locus 1	0.083	0.046	0.26	0.34	0.32
Locus 2	0.083	0.49	0.26	0.34	0.32
Epistasis	0.83	0.47	0.48	0.32	0.37

The models are a selection of those explored in Evans et al. [2006]: recessive-recessive (RR), recessive-dominant (RD), dominant-dominant (DD) and dominant-dominant, except that the double heterozygote does not have the effect (DRD) and multiplicative interaction (M).

interaction. Model M is most similar to models DD and DRD, hence we expect SC to be most challenged by the recessive and partially recessive models. In our simulations, as expected, power is similar to model M for models DD and DRD, but not promising for models RR and RD.

Next we try a simulation of case and control data. We simulate 600 cases and 600 controls from a population of SNPs with first-order Markov dependency $\rho = 0.3$, MAF = 0.3. The data were generated using the following model with five pair-wise interactions of randomly selected SNPs.

$$\text{logit} = \beta(X_1X_2 + 1.5X_3X_4 + 2X_5X_6 + 2.5X_7X_8 + 3X_9X_{10}) - 2.$$

We simulated two types of data sets, one with 5,000 SNPs and the other with 50,000 SNPs. In the situation of 50,000 SNPs, we first select top 5,000 SNPs by marginal test using logistic regression. Then, for both types of the data sets, we apply the SC_i procedure. The power was essentially equivalent for both scenarios. The Type I error increased from approximately 0.05 to 0.15.

ANALYSIS OF T1D DATA

The Wellcome Trust Case Control Consortium [WTCCC, 2007] data includes 1,963 cases with T1D and 2,938 controls (post QC) obtained from people living in Great Britain who self-identified as white Europeans; see WTCCC [2007] for details about the sample and QC procedures. The samples were genotyped with the GeneChip 500K Mapping Array Set (Affymetric chip).

At least 12 regions in the genome have strong statistical support in the literature for association with T1D (i.e., 1p13.2, PTPN22; 2p24.2, IFIH1; 2q33.2, CTLA4; 6p21.32, HLA class II; 6q15, BACH2; 10p15.1, PRKCQ; 11p15.5, INS; 12q12.2, ERBB3; 12q24.13, SH2B3/PTPN11; 15q25.1, CTSH; 16q13.13, CLEC16A; 18p11.21 PTPN2) with reported p-values less than 10^{-8} [Hindorff et al., 2009]. Univariate analyses of the WTCCC data show genome-wide significance for four of these (i.e., PTPN22, HLA class II, ERBB3, and the SH2B3/PTPN11 region). In addition rs12708716 (CLEC16A) on Chromosome 16 is borderline significant. The INS gene, which is not tagged well by this array, does not show evidence of association in these data [WTCCC, 2007].

We reanalyzed these data using the multi-split SC approach with 56 random splits of the data (1/3 screen and 2/3 clean). Of the 469,612 SNPs passing WTCCC QC, we also removed 594 SNPs with poor genotype clustering patterns and all SNPs on chromosome X. Next we restricted the dictionary to those with univariate p-value less than 0.017. From the remaining 10,000 SNPS, we chose SNPs using H-clust, set to pick tag SNPs with squared correlation less than 0.04 and MAF greater than 0.01; for a cluster of SNPs in LD, H-clust used preference scores based on the univariate p-values for association of each SNP [Rinaldo et al., 2005]. In this way, our tag SNP selection process includes the SNP most likely to be associated with T1D within each LD block. After this process, we further ensured that the SNP dictionary included no SNPs with squared correlation greater than 0.045 on the same chromosome. The resulting dictionary included 3,437 SNPs. We recoded the genotype data for the additive model. For the SC_i algorithm, to keep the model size computationally manageable we used $R_1 = 250$ and $R_2 = 2000$.

TABLE III. Best SNP main effects found via SC_m and corrected for multiple testing

Chr	Position (bp)	SNP	SC p-value	Univariate p-value
1	114105331	rs6679677	5.6×10^{-14}	5.1×10^{-25}
4	123548812	rs17388568	0.35	5.7×10^{-7}
6	31735428	rs2242655	1.8×10^{-2}	5.4×10^{-6}
6	32297010	rs415929	1.8×10^{-2}	2.9×10^{-5}
6	32712350	rs9272346	1.1×10^{-76}	8.9×10^{-122}
6	32910181	rs241432	3.5×10^{-4}	1.7×10^{-6}
6	33111665	rs448733	2.7×10^{-2}	1.1×10^{-5}
12	54756892	rs11171739	2.6×10^{-5}	1.3×10^{-11}
12	110971201	rs17696736	1.3×10^{-2}	1.0×10^{-11}
16	11115395	rs9746695	1.4×10^{-3}	9.6×10^{-9}

Nominal univariate p-values, not corrected for multiple testing, are obtained using logistic regression.

Our results are similar to the WTCCC's univariate analysis (Table III). All five of their best signals also appeared as significant effects in our model. In addition, on Chromosome 4, SNP rs17388568, which was borderline significant (5.7×10^{-7}) using conditional logistic regression, is also borderline in our analysis (multiple testing corrected p-value = 0.35). Our model also identified four additional SNPs in the HLA region. Because we restricted our analysis to tag SNPs the LD between these SNPs is minor, suggesting the signal in the MHC is due to multiple variants.

Applying SC_i we identified three SNP-SNP interactions as significant (Table IV), corresponding to univariate SNP-SNP p-values that would not have been sufficient to attain genome-wide significance in a standard analysis. This suggests that the SC_i procedure can indeed be more powerful than a series of univariate tests, especially when searching through the vast model space of SNP-SNP interactions.

Two of the pairs involve SNPs in the MHC region. Both of these include an SNP that was identified as main effects paired with another that did not demonstrate significant main effects (rs241429, univariate p-value of 8.2×10^{-5}). The remaining interaction involves a pair of SNPs on Chromosome 12, one discovered as a main effect (rs17696736) that tags the SH2B3/PTPN11 region, paired with (rs11066119, univariate p-value of 9.6×10^{-5}). Pairs of SNPs are not in linkage disequilibrium (Table IV).

TABLE IV. Best SNP interaction effects found via SC_i and corrected for multiple testing

						SC	Univariate		
Chr	Position (bp)	SNP	Chr	Position (bp)	SNP	*p*-value	*p*-value	*p*-value	r^2
6	32911818	rs241429	6	32910181	rs241432	3.5×10^{-66}	8.2×10^{-5}	1.7×10^{-6}	0.02
6	32911818	rs241429	6	32712350	rs9272346	1.9×10^{-25}	8.2×10^{-5}	9.0×10^{-122}	<0.001
12	110918887	rs11066119	12	110971201	rs17696736	2.3×10^{-15}	9.6×10^{-5}	1.0×10^{-11}	<0.001

Nominal univariate *p*-values, not corrected for multiple testing, are obtained using logistic regression.

Moreover, because each of these variants is significant in the multivariate model, we can conclude that each variant exhibits a significant association with the phenotype, after accounting for all of the other variants in the model. The genotype by genotype counts support our findings (Supplementary Table A).

Our initial analyses of these data, conducted after removal of SNPs that failed standard QC measures but prior to removal of SNPs that failed visual QC inspection, identified several additional SNP-SNP interactions (results not shown). Unfortunately, these SNPs did not pass the visual QC inspection performed by WTCCC. From this we conclude that interactions are much more sensitive to poor genotype quality than tests of main effects, and care must be taken to thoroughly inspect the data for problems with genotype calling. Our reported results were conducted after removal of all SNPs with small univariate *p*-values that showed poor genotype clustering.

DISCUSSION

A multivariate regression model can have greater power than a series of marginal tests to detect signals when multiple variables affect the outcome, as seen here and in other research [Chatterjee et al., 2006; Hoggart et al., 2008; Longmate, 2001; Millstein et al., 2006; Ritchie et al., 2003; Zhang and Liu, 2007]. Building on this idea we propose the SC algorithm, which identifies the most promising SNPs and interactions simultaneously using the lasso regression procedure. Because the class of models that includes all potential interactions is too large to be practical, we vary the dictionary of SNPs and SNP-SNP interactions considered at each step of the analysis. First only main effects are considered. Next we include interactions corresponding to SNPs that exhibit at least a weak main effect. Using an independent source of data, in the final step we look for replication of those terms that look most promising in the first step analysis. This approach lies somewhere between classical replication analysis and joint analysis of the data [Skol et al., 2006]. Contrary to joint analyses, only the replication data are used in the validation study, but SNPs and SNP-SNP interactions that go on to the second stage for validation need not exceed the threshold for genome-wide significance in stage one.

We applied our procedure on data simulated with linkage disequilibrium and data similar to a GWAS. Because SC is designed to model the effects of a multiple gene system it provides good control of the type I error and FDR even when the SNPs are in LD. Although many marginal tests are available in the literature, we compared our results with two methods that work well with data collected in two stages using a joint analysis approach [Kooperberg and LeBlanc, 2008; Lin, 2006]. In data simulated to mimic a GWAS with many SNP and SNP-SNP interactions present, the marginal methods lacked power relative to the SC approach. It is not surprising that a multiple regression approach has greater potential to handle these complex models than a sequential approach based on marginal tests because it reduces the variance and allows the causal SNPs, or SNPs highly correlated with causal SNPs, to compete for variance prediction against other SNPs having no impact.

From statistical theory and practice we know that regression models that include highly correlated predictors have myriad undesirable properties. Consistent with the theory our simulation results show that SC works better when linkage disequilibrium among SNPs is small. Moreover, including correlated SNPs increases the computational burden severely without adding substantial information. Thus, we recommend always using tag SNPs for SC. To enhance the chance that associated SNPs are included in the set of tag SNPs we suggest choosing the SNP with the smallest marginal *p*-value among any correlated set of SNPs [Rinaldo et al., 2005]. After applying the SC procedure to the tag SNPs, one can investigate all of the SNPs genotyped in the vicinity of the tag SNPs identified in the initial analyses.

Contrary to some methods in the literature, SC identifies promising main effects, followed by pursuit of epistatic effects. Of course, some kinds of epistasis do not lend themselves to a two-step approach because the majority of the genetic variance resides strictly in the interactions and hence the SNPs cannot be identified via single-locus tests of association [Evans et al., 2006]. For other epistatic models, however, the power to detect each locus using a single-locus strategy is high enough that a two-step strategy does have advantages. When using the lasso, success in finding individual SNPs in step 1 is further enhanced because the model space is multivariate. Nevertheless, if a model has low variance attributable to either of the two SNPs involved, there is little chance that the epistatic effect will be discovered with a two-step approach.

Our approach differed from that of Evans et al. in a number of ways. Regardless of the true model, we used an additive model for the SC. With this approach we gain power when the model is approximately additive by using fewer degrees of freedom, but we lose power when the model is far from additive. In contrast, Evans et al. use a genotype model that allows for multiple degrees of freedom for single-locus and two-locus tests. Our simulations were designed for data collected from a two-stage experimental design. With this design it is assumed that only the most promising SNPs are evaluated at stage 2.

Thus, SC is not limited to an additive model, but can be used for any genetic model or family of models. With SC the second set of data is utilized to clean the model of superfluous terms. The other approaches include an implicit correction for joint analysis of data from both stages of data. This correction is more powerful than the one utilized by Evans et al. in their simulations.

An option, which we did not pursue in this paper, is SC applied directly to a dictionary including all main effects and interactions without the benefit of screening. This approach is not computationally feasible if the number of SNPs is large. To bypass the hierarchical search and yet avoid severe computational hurdles one could combine the best features of marginal testing and the multivariate approach: define the preliminary dictionary to be the set of all possible pairwise interactions; contract this huge dictionary by testing for main effects and interactions using a marginal test to obtain a dictionary corresponding to terms with smallest p-values; and screen this dictionary using stage 1 data and clean using stage 2 data. This approach is not limited by the assumed hierarchical structure, but it has the disadvantage of being very computationally intensive [Marchini et al., 2005; Purcell et al., 2007].

In principle we would like to create even richer interaction dictionaries. One way to consider more SNPs, and yet achieve the advantages of the multivariate analysis, is to split the SNPs into subcategories that are more likely to be involved in interactions. For instance, we might choose subsets of SNPs from different pathways and create pathway-dependent dictionaries [Bochdanovits et al., 2008; Emily et al., 2009]. With this approach, each dictionary is less likely to exceed the computational limit, and yet likely interactions are included in the screening process. This approach could be successful in discovering complex epistatic models.

Application of the method to data obtained from the Wellcome Trust Case Control Consortium study of T1D cases and controls uncovered evidence supporting multiple HLA class II independent T1D associations within the HLA class I regions occurring at HLA-B and HLA-A [Howson et al., 2009; Nejentsev et al., 2007; Valdes et al., 2005]. Analyses of interacting SNP pairs discovered association occurring within HLA class II as well as within the Chromosome 12q24 region. The HLA region represents the largest genetic risk element for T1D as well as other autoimmune diseases [Klein and Sato, 2000]. A likely mechanism by which certain HLA alleles influence T1D susceptibility is related to their ability to bind and present autoantigens to autoreactive T-lymphocytes in the thymus [Morel et al., 1988; Nepom and Erlich, 1991; Todd et al., 1987]. Likewise, the Chromosome 12q24 region has been confirmed as associated with T1D [Barrett et al., 2009; Todd et al., 2007]. These studies have identified a large LD block, estimated at greater than 1.2 Mb, harboring at least two genes with possible functional relevance to T1D, such as PTPN11 and SH2B3 [Smyth et al., 2008; Todd et al., 2007].

SNP pairs that interact to influence disease susceptibility may do so by mechanisms involving transcriptional control, mRNA processing, changes in amino acid sequence, or a combination of mechanisms. For example, HLA loci polymorphisms in the promoter region have been described that result in changes in expression [Beaty et al., 1995]. Non-HLA loci that influence T1D risk have also been linked to changes in expression (i.e., INS) as well as altered RNA processing (i.e., CTLA4) [Ounissi-Benkalha and Polychronakos, 2008]. In these examples altered expression has been proposed to affect autoimmunity by influencing negative selection of autoreactive T-lymphocytes [Fan et al., 2009]. The SNP pairs identified by our analyses may also impact gene expression; however, additional experiments will be needed to sufficiently characterize these elements in order to elucidate their mechanism of interaction with T1D risk.

ACKNOWLEDGMENTS

This work was funded by the National Institutes of Health grant MH057881 awarded to B. D. and K. R. and by the Department of Defense (grant W81XWH-07-1-0619) awarded to M. T. This study makes use of data generated by the Wellcome Trust Case-Control Consortium 2. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk ⟨http://www.wtccc.org.uk/⟩. Funding for the project was provided by the Wellcome Trust under award 085475. Electronic References: The R code is available at http://www.wpic.pitt.edu/WPICCompGen/.

REFERENCES

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Plagnol V, Pociot F, Schuilenburg H, Smyth DJ, Stevens H, Todd JA, Walker NM, Rich SS, The Type 1 Diabetes Genetics Consortium. 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41:703–707.

Beaty JS, West KA, Nepom GT. 1995. Functional effects of a natural polymorphism in the transcriptional regulatory sequence of hla-dqb1. Mol Cell Biol 15:4771–4782.

Bochdanovits Z, Sondervan D, Perillous S, van Beijsterveldt T, Boomsma D, Heutink P. 2008. Genome-wide prediction of functional gene-gene interactions inferred from patterns of genetic differentiation in mice and men. PLoS ONE 3:e1593.

Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. 2006. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet 79:1002–1016.

Cordell HJ. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11:2463–2468.

Cordell HJ. 2009. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10:392–404.

de Bakker PIW, Yelensky R, Peér I, Gabriel SB, Daly JJ, Altshuler D. 2005. Efficiency and power in genetic association studies. Nat Genet 37:1217–1223.

Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322.

Devlin B, Roeder K, Wasserman L. 2003. Analysis of multilocus models of associaiton. Genet Epidemiol 25:36–47.

Emily M, Mailund T, Schauser L, Schierup MH. 2009. Using biological networks to search for interacting loci in genomewide association studies. Eur J Hum Genet 17:1231–1240.

Evans DM, Marchini J, Morris AP, Cardon LR. 2006. Two-stage two-locus models in genome-wide association. PLoS Genet 2:e157.

Fan Y, Rudert WA, Grupillo M, He J, Sisino G, Trucco M. 2009. Thymus-specific deletion of insulin induces autoimmune diabetes. EMBO J 00:00–00.

Hastie T, Tibshirani R, Friedman J. 2001. The Elements of Statistical Learning. New York: Springer.

Hindorff LA, Junkins HA, Mehta JP, Manolio TA. 2009. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/26525384.

Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. 2008. Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. PLoS Genet 4:e1000130.

Hoh J, Wille A, Zee R, Cheng S, Reynolds R, Lindpaintner K, Ott J. 2000. Selecting snps in two-stage analysis of disease association data: a model-free approach. Ann Hum Genet 64:413–417.

Howson JM, Walker NM, Clayton D, Todd JA, Diabetes Genetics Consortium. 2009. Confirmation of hla class ii independent type 1 diabetes associations in the major histocompatibility complex including hla-b and hla-a. Diabetes Obes Metab 11:31–45.

Klein J, Sato A. 2000. The hla system. N Engl J Med 343:782–786. Second of two parts.

Kooperberg C, LeBlanc M. 2008. Increasing the power of identifying gene × gene interactions in genome-wide association studies. Genet Epidemiol 32:255–263.

Lee AB, Luca D, Klei L, Devlin B, Roeder K. 2010. Discovering genetic ancestry using spectral graph theory. Genet Epidemiol 34:51–59.

Lin DY. 2006. Evaluating statistical significance in two-stage genome-wide association studies. Am J Hum Genet 78:505–509.

Longmate JA. 2001. Complexity and power in case-control association studies. Am J Hum Genet 68:1229–1237.

Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M. 2008. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82:453–463.

Manolio TA, Collins F. 2007. Genes, environment, health, and disease: facing up to complexity. Hum Hered 63:63–66.

Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci influencing complex diseases. Nat Genet 37:413–417.

Meinshausen N, Meier L, Buhlmann P. 2008. *P*-values for high dimensional regression. arXiv: 0811.2177v2.

Millstein J, Conti DV, Gilliland FD, Gauderman WJ. 2006. A testing framework for identifying susceptibility genes in the presence of epistasis. Am J Hum Genet 78:15–27.

Morel PA, Dorman JS, Todd JA, McDevitt HO, Trucco M. 1988. Aspartic acid at position 57 of the hla-dq beta chain protects against type i diabetes: a family study. Proc Natl Acad Sci USA 85:8111–8115.

Nejentsev S, Howson JM, M WN, Szeszko J, Field SF, Stevens HE, Reynolds P, Hardy M, King E, Masters J, Hulme J, Maier LM, Smyth D, Bailey R, Cooper JD, Ribas G, Campbell RD, Clayton DG, Todd JA, Wellcome Trust Case Control Consortium. 2007. Localization of type 1 diabetes susceptibility to the mhc class i genes hla-b and hla-a. Nature 450:887–892.

Nepom GT, Erlich H. 1991. Mhc class-ii molecules and autoimmunity. Annu Rev Immunol 9:493–525.

Ounissi-Benkalha H, Polychronakos C. 2008. The molecular genetics of type 1 diabetes: new genes and emerging mechanisms. Trends Mol Med 14:268–275.

Phillips PC. 2008. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9:855–867.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, de Bakker PIW, Daly MJ, Sham PC. 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575.

Rinaldo A, Bacanu SA, Devlin B, Sonpar V, Wasserman L, Roeder K. 2005. Characterization of multilocus linkage disequilibrium. Genet Epiodemiol 28:193–206.

Ritchie MD, Hahn LW, Moore JH. 2003. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epiodemiol 24:150–157.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213.

Smyth DJ, Cooper JD, Howson JM, Walker NM, Plagnol V, Stevens H, Clayton DG, Todd JA. 2008. Ptpn22 trp620 explains the association of chromosome 1p13 with type 1 diabetes and shows a statistical interaction with hla class ii genotypes. Diabetes 57:1730–1737.

Strobl C, Boulesteix AL, Zeileis A, Hothorn T. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8:25.

The Welcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 controls. Nature 447:661–678.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. J Roy Stat Soc B 58:267–288.

Todd JA, Bell JI, McDevitt HO. 1987. Hla-dq beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. Nature 329:599–604.

Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, Lowe CE, Szeszko JS, Hafler JP, Zeitels L, Yang JH, Vella A, Nutland S, Stevens HE, Schuilenburg H, Coleman G, Maisuria M, Meadows W, Smink LJ, Healy B, Burren OS, Lam AA, Ovington NR, Allen J, Adlem E, Leung HT, Wallace C, Howson JM, Guja C, Ionescu-Tirgovirte C, Genetics of Type 1 Diabetes in Finland, Simmonds MJ, Heward JM, Gough SC, Wellcome Trust Case Control Consortium, Dunger DB, Wicker LS, Clayton DG. 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39:857–864.

Valdes AM, Erlich HA, Noble JA. 2005. Human leukocyte antigen class i b and c loci contribute to type 1 diabetes (t1d) susceptibility and age at t1d onset. Hum Immunol 66:301–313.

Wasserman L, Roeder K. 2009. High dimensional variable selection. Annal Stat 37:2178–2201.

Wu TT, Chen YF, Hastie T, Sobel E, Lange K. 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25:714–721.

Zhang Y, Liu JS. 2007. Bayesian inference of epistatic interactions in case-control studies. Nat Genet 39:1167–1173.

FoxO1 Links Hepatic Insulin Action to Endoplasmic Reticulum Stress

Adama Kamagate,* Dae Hyun Kim,* Ting Zhang, Sandra Slusher, Roberto Gramignoli, Stephen C. Strom, Suzanne Bertera, Steven Ringquist, and H. Henry Dong

Division of Immunogenetics (A.K., D.H.K., T.Z., S.S., S.B., S.R., H.H.D.), Rangos Research Center, Children's Hospital of Pittsburgh of University of Pittsburgh Medical Center, Department of Pediatrics, Pittsburgh, Pennsylvania 15224; and Department of Pathology (R.G., S.C.S.), University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213

Forkhead box O1 (FoxO1) is a transcription factor that mediates the inhibitory effect of insulin on target genes in hepatic metabolism. Hepatic FoxO1 activity is up-regulated to promote glucose production during fasting and is suppressed to limit postprandial glucose excursion after meals. Increased FoxO1 activity augments the expression of insulin receptor (IR) and IR substrate (IRS)2, which in turn inhibits FoxO1 activity in response to reduced insulin action. To address the underlying physiology of such a feedback loop for regulating FoxO1 activity, we delivered FoxO1-ADA by adenovirus-mediated gene transfer into livers of adult mice. FoxO1-ADA is a constitutively active allele that is refractory to insulin inhibition, allowing us to determine the metabolic effect of a dislodged FoxO1 feedback loop in mice. We show that hepatic FoxO1-ADA production resulted in significant induction of IR and IRS2 expression. Mice with increased FoxO1-ADA production exhibited near glycogen depletion. Unexpectedly, hepatic FoxO1-ADA production elicited a profound unfolded protein response, culminating in the induction of hepatic glucose-regulated protein 78 (GRP78) expression. These findings were recapitulated in primary human and mouse hepatocytes. FoxO1 targeted *GRP78* gene for *trans*-activation via selective binding to an insulin responsive element in the GRP78 promoter. This effect was counteracted by insulin. Our studies underscore the importance of an IR and IRS2-dependent feedback loop to keep FoxO1 activity in check for maintaining hepatic glycogen homeostasis and promoting adaptive unfolded protein response in response to altered metabolism and insulin action. Excessive FoxO1 activity, resulting from a dislodged FoxO1 feedback loop in insulin resistant liver, is attributable to hepatic endoplasmic reticulum stress and metabolic abnormalities in diabetes. (*Endocrinology* 151: 3521–3535, 2010)

Forkhead box O1 (FoxO1) belongs to a superfamily of transcription factors that is characterized by a highly conserved winged-helix DNA binding motif, termed "forkhead" domain, including FoxO1, FoxO3, FoxO4, and FoxO6 in mammals (1, 2), abnormal dauer formation (DAF) 16 in *Caenorhabditis elegans* (3), and dFoxO in *Drosophila* (4). These forkhead proteins are substrates of serine-threonine kinase/protein kinase B and serum/glucocorticoid regulated kinase, playing important roles in mediating insulin action on the expression of genes involved in cell growth, differentiation, metabolism, and longevity (1–5). Insulin exerts its inhibitory effect on target gene expression via a highly conserved insulin responsive element (IRE) with its core motif 5′-TG/ATTTT/G-3′

ISSN Print 0013-7227 ISSN Online 1945-7170
Printed in U.S.A.

doi: 10.1210/en.2009-1306 Received November 5, 2009. Accepted April 29, 2010.
First Published Online May 25, 2010
* A.D. and D.H.K. contributed equally to this work.

Abbreviations: ATF, Activating transcription factor; ChIP, chromatin immunoprecipitation; CHOP, CCAAT-enhancer-binding protein homology protein; ER, endoplasmic reticulum; FFA, free fatty acid; FoxO1, forkhead box O1; GRP78, glucose-regulated protein 78; IR, insulin receptor; IRS, IR substrate; IRE, insulin responsive element; IRE1, inositol requiring 1; MOI, multiplicity of infection; nt, nucleotide; PBA, 4-phenyl butyric acid; PERK, protein kinase R-like ER kinase; pfu, plaque-forming unit; siRNA, small interfering RNA; UPR, unfolded protein response; VLDL, very low-density lipoprotein.

in the promoter (1, 2, 6). In response to reduced insulin action, FoxO proteins reside in the nucleus and bind as a *trans*-activator to IRE, enhancing promoter activity. In response to insulin stimulation, FoxO proteins are phosphorylated through the phosphatidylinositol kinase-dependent pathway, resulting in FoxO nuclear exclusion and inhibition of target gene expression (7–13). This phosphorylation-dependent subcellular redistribution serves as an acute mechanism for insulin to regulate FoxO transcriptional activity for rapid adaptation to metabolic shift from fasting to refeeding states (1, 2, 5, 6, 14). Except for FoxO6 (15), all members of the FoxO superfamily undergo insulin-dependent phosphorylation and nuclear exclusion. Failure in FoxO phosphorylation results in its permanent nuclear localization and constitutive gene expression (1, 2, 6, 7, 16–18). Indeed, it has been shown that unbridled FoxO1 activity, resulting from an impaired ability of insulin to phosphorylate FoxO1, promotes the overproduction of gluconeogenic enzymes PEPCK and G6PC (1, 5, 19–23), as well as apolipoprotein C-III and microsomal triglyceride transfer protein, two key functions in very low-density lipoprotein (VLDL)-triglyceride metabolism (24–26). This effect accounts in part for the concurrent pathogenesis of fasting hyperglycemia and hypertriglyceridemia in insulin-resistant subjects with visceral obesity and type 2 diabetes (6, 27, 28).

There is anecdotic evidence that FoxO1 activity is subject to feedback regulation, but the underlying physiology remains elusive. FoxO1 is shown to stimulate the expression of its upstream effector gene encoding insulin receptor (IR), which in turn activates insulin signaling and inhibits FoxO1 activity (29, 30). It is postulated that such a feedback loop serves as a mechanism for enhancing cellular sensitivity to insulin during fasting and priming starved cells for nutrient availability. Implicit in this assumption is that FoxO1 activity is up-regulated in serum-starved cells (29, 30) and in liver of fasted mice (17). However, this hypothesis seems at odds with the clinical data showing that prolonged fasting (60 h) elicits peripheral insulin resistance with a concomitant induction in plasma free fatty acid (FFA) levels in healthy subjects (31). A 16-h fast results in increased lipid accumulation in liver without affecting insulin sensitivity in mice (32). Excessive FoxO1 activity also results in hepatosteatosis in mice (5, 17, 21, 26, 33, 34).

To address the physiological significance of the FoxO1 feedback loop, we delivered FoxO1-ADA by adenovirus-mediated gene transfer into liver of adult mice. FoxO1-ADA is a constitutively active allele that is not subject to insulin inhibition due to point mutations in the three conserved phosphorylation sites (T24A, S253D, and S316A) of FoxO1 polypeptide chain (10, 19). As a result, this system would disengage the effect of insulin on FoxO1 activity, allowing us to determine the metabolic consequence of a dislodged FoxO1 feedback loop in adult mice. We show that hepatic FoxO1-ADA production resulted in a significant induction of IR and IR substrate (IRS)2. Unexpectedly, hepatic FoxO1-ADA production selectively enhanced the expression of glucose-regulated protein 78 (GRP78), a molecular chaperone that resides in the endoplasmic reticulum (ER) and functions as an ER stress sensor to maintain ER homeostasis (35–37). This effect correlated with near depletion of hepatic glycogen content in mice with elevated FoxO1-ADA production. We recapitulated these findings in cultured hepatocytes with elevated FoxO1-ADA production. Furthermore, we show that FoxO1 stimulated GRP78 promoter activity via specific binding to its consensus IRE motif within the GRP78 promoter. This effect was counteracted by insulin. Mutations or deletion of the IRE motif resulted in abolition of FoxO1-mediated induction of GRP78 expression. In addition, we show that palmitate, a predominant saturated form of FFA that is known to elicit ER stress, augmented hepatic FoxO1 activity and induced GRP78 production. Palmitate-mediated induction of FoxO1 and GRP78 production was reversed to normal in response to 4-phenyl butyric acid (PBA), a pharmacological chaperone that is effective for mitigating cellular ER stress (38). Moreover, enhanced binding of FoxO1 to GRP78 promoter was detectable in insulin resistant liver, correlating with augmented hepatic FoxO1 activity and increased GRP78 production in obese *db/db* mice.

These results characterize GRP78 as a molecular target of FoxO1, underscoring the importance of FoxO1 in hepatic ER homeostasis. ER is the principal organelle for the biosynthesis of proteins and steroids and for the production of VLDL particles. Perturbation of ER homeostasis, such as the accumulation of misfolded proteins, deprivation of glucose, or altered glycosylation, often triggers adaptive unfolded protein response (UPR), also known as ER stress (35–37, 39–42). Unresolved ER stress results in cellular apoptosis (36, 43–47). Although UPR is intertwined with impaired insulin action and there is emerging evidence that excessive ER stress is attributable to insulin resistance (35, 48, 49), the underlying mechanism remains elusive. Our results together with previous data indicate that FoxO1 integrates hepatic insulin action to GRP78 expression for regulating UPR in a pathway that is orchestrated through the IR/IRS2-dependent FoxO1 feedback loop. We suggest that the FoxO1 feedback loop is crucial for keeping FoxO1 activity in check, a safeguarding mechanism for maintaining ER homeostasis and averting the deleterious effect of unrestrained FoxO1 activity on glucose and lipid metabolism.

Materials and Methods

Animal studies

CD1 mice were obtained from Charles River Laboratory (Wilmington, MA). For blood chemistry, mice were fasted for 16 h, and tail vein blood samples were collected into capillary tubes precoated with potassium-EDTA (Sarstedt, Nümbrecht, Germany) for the preparation of plasma or determination of blood glucose levels using Glucometer Elite (Bayer, Mishawaka, IN) and plasma insulin using the ultrasensitive mouse insulin ELISA (ALPCO, Windham, NH). All procedures were approved by the Institutional Animal Care and Use Committee of University of Pittsburgh School of Medicine. Other methods, including statistics, were described in online Supplemental Materials and Methods (published on The Endocrine Society's Journals Online web site at http://endo.endojournals.org).

Results

FoxO1 up-regulates hepatic IR and IRS2 expression in mice

FoxO1 acts downstream of IR and IRS to mediate insulin action on hepatic gluconeogenesis. To investigate the effect of FoxO1 on IR and IRS expression, we delivered the constitutively active FoxO1-ADA allele into liver of mice using adenovirus-mediated gene transfer. Due to amino acid substitutions at the three conserved phosphorylation sites, FoxO1-ADA is refractory to insulin inhibition (19, 26), allowing us to determine the net effect of FoxO1 on the expression of its upstream effectors, such as IR and IRS in liver. Male CD1 mice were stratified by body weight into two groups (n = 8 per group), which were treated with FoxO1-ADA and control vectors, as described (25). As shown in Fig. 1, hepatic FoxO1-ADA production resulted in a 2-fold induction in both IR and IRS2 protein levels. This effect correlated with a 2.5-fold elevation in hepatic FoxO1 activity in FoxO1-ADA group. In contrast, hepatic expression of IRS1 proteins remained unchanged in FoxO1-ADA *vs.* control vector-treated mice. These data suggest that FoxO1-ADA selectively up-regulated hepatic IR and IRS2 expression in mice.

FoxO1 stimulates IR and IRS2 expression in cultured hepatocytes

To corroborate the above findings, we treated HepG2 cells in the presence and absence of FoxO1-ADA production, followed by the determination of IR and IRS expression. FoxO1-ADA production resulted in more than 4-fold induction in both IR (Fig. 1E) and IRS2 (Fig. 1F) protein levels, correlating with a 6-fold elevation of FoxO1 activity in FoxO1-ADA vector-treated HepG2 cells (Fig. 1G). Due to extremely low basal IRS1 expression, we could not detect IRS1 protein expression in HepG2 cells. Although FoxO1 activity is inhibited by insulin via an IR- and IRS-dependent mechanism, the present findings of FoxO1-mediated induction of IR and IRS2 are consistent with the idea that hepatic FoxO1 activity is tightly regulated via a feedback loop (30).

Correlation of FoxO1 activity with IR and IRS2 expression in fasted liver

To understand the underlying physiology of the FoxO1 feedback loop, we determined the expression of hepatic IR, IRS1, and IRS2 in mice under different physiological conditions. Based on previous observations that FoxO1 protein expression along with its nuclear localization is increased, correlating with its enhanced activity in promoting gluconeogenesis in response to fasting (17), we hypothesized that increased FoxO1 activity would stimulate IR and IRS2 expression in fasted livers. Two groups of male CD1 mice (n = 6 per group) were maintained under fed conditions or fasted for 16 h, followed by determining IR and IRS protein levels in liver. As shown in Fig. 2, IR and IRS2 protein levels were significantly raised when the metabolic state was shifted from fed to an overnight fasting. This effect correlated with elevated hepatic FoxO1 activity, culminating in significantly increased FoxO1 nuclear localization in hepatocytes of fasted mice, in accordance with our previous observations (17). A small nonsignificant increase in hepatic IRS1 protein levels was detected in fasted mice.

Effect of FoxO1 on hepatic glycogen metabolism

To address the underlying pathophysiology of FoxO1 feedback loop, we probed the metabolic consequence of unbridled FoxO1 activity in liver, resulting from a lack of FoxO1 feedback inhibition. We determined hepatic glycogen content in FoxO1-ADA vector-treated mice. Due to its insensitivity to insulin inhibition, FoxO1-ADA production provides a scenario of a circuit breakdown in FoxO1 feedback loop. As a result, mice with increased FoxO1-ADA production in liver displayed a 3.5-fold reduction in hepatic glycogen content after an overnight fasting (Fig. 3A). Hepatic glycogen was nearly depleted (<10 mg/g liver protein) in three out of eight FoxO1-ADA vector-treated mice, in comparison with hepatic glycogen content (>70 mg/g liver protein) in control vector-treated mice. This effect correlated with the ability of FoxO1 to stimulate G6PC production and promotes gluconeogenesis and glycogenolysis in liver (1, 23, 26). No significant differences in body weight were detected after 1 wk of hepatic FoxO1-ADA production, ruling out the possibility that the observed alterations in hepatic glycogen metabolism were secondary to body weight changes in FoxO1-ADA vector-treated mice.

Furthermore, FoxO1-ADA vector-treated mice exhibited relatively lower blood glucose levels (Fig. 3B). This

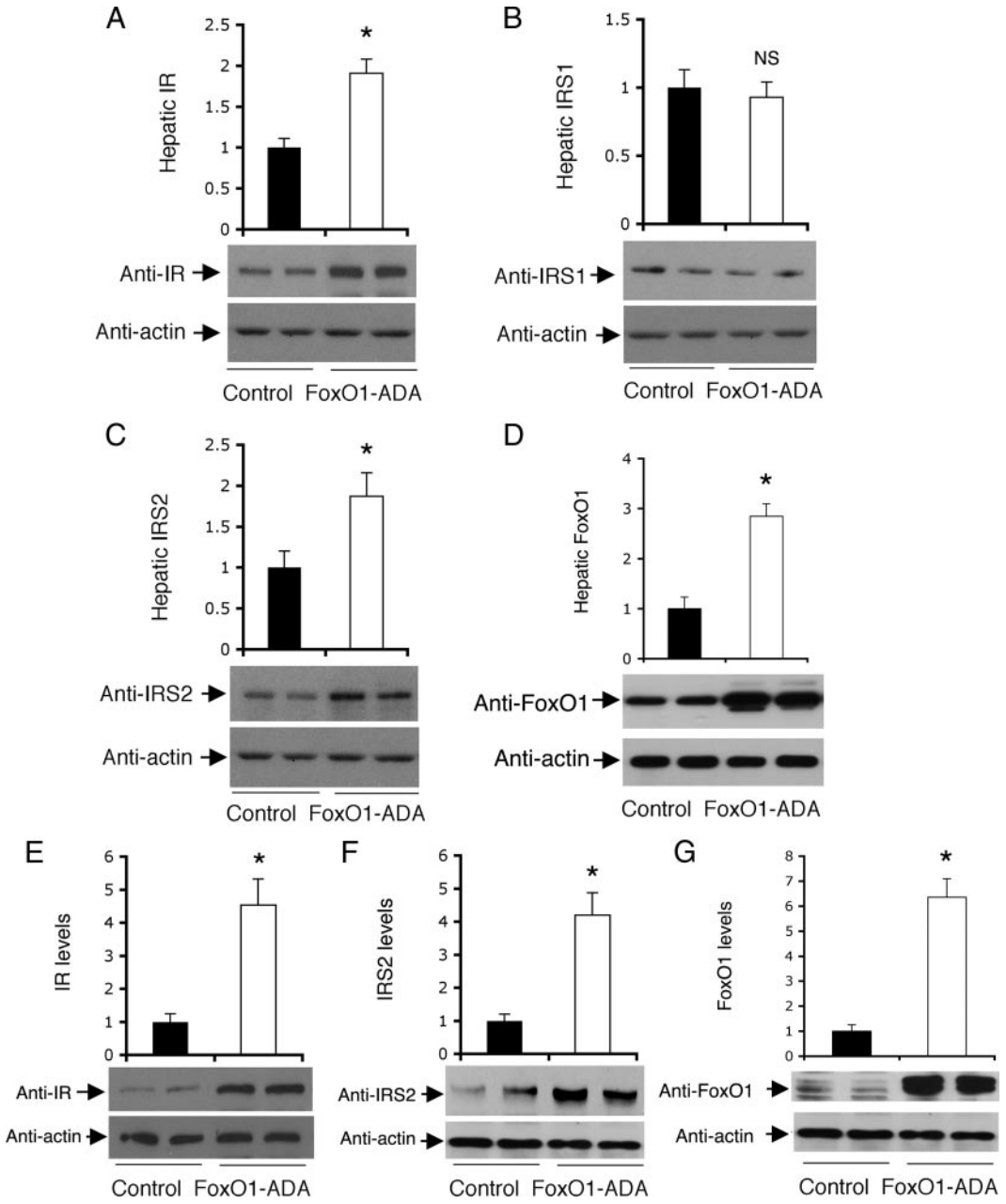


FIG. 1. Effect of FoxO1-ADA on hepatic IR and IRS expression. Male CD1 mice of 10 wk old were stratified by body weight and randomly assigned to two groups (n = 8), which were iv injected with Adv-FoxO1-ADA or Adv-null vector at 1.5×10^{11} pfu/kg body weight. One week after vector administration, mice were fasted for 16 h and killed. Liver tissues were subjected to immunoblot analysis for the determination of hepatic levels of IR (A), IRS1 (B), IRS2 (C), and FoxO1 (D). HepG2 cells at 90% confluence were transduced with Adv-FoxO1-ADA or Adv-null vector at a fixed MOI (100 pfu/cell) in six-well plates. Each condition was run in triplicate. After a 24-h incubation, cells were subjected to immunoblot analysis using anti-IR (E), anti-IRS2 (F), and FoxO1 (G). IRS1 was undetectable due to its extremely low basal expression in HepG2 cells. *, $P < 0.05$ *vs.* control. NS, Not significant.

effect was inversely correlated with significantly increased fasting plasma insulin levels (Fig. 3C). Similar effects have been observed in adult mice with increased FoxO1 activity in liver in previous studies (17, 26).

To underpin the above findings, we cultured human primary hepatocytes in 12-well collagen-coated microplates at the density of 1×10^6 cells/well in the presence of FoxO1-ADA or control vector at a fixed dose [multiplicity of infection (MOI), 200 plaque-forming unit (pfu)/cell]. After a 16-h incubation, cells were harvested for the determination of intracellular glycogen. Adenovirus-mediated FoxO1-ADA production resulted in a 5-fold reduction of glycogen content in human primary hepatocytes (Fig. 3D). Likewise, we treated mouse primary hepatocytes with 200 MOI of Adv-null vector or Adv-FoxO1-ADA, followed by the determination of hepatic glycogen. Consistent with the observation in human primary hepatocytes, mouse primary hepatocytes with FoxO1 gain-of-

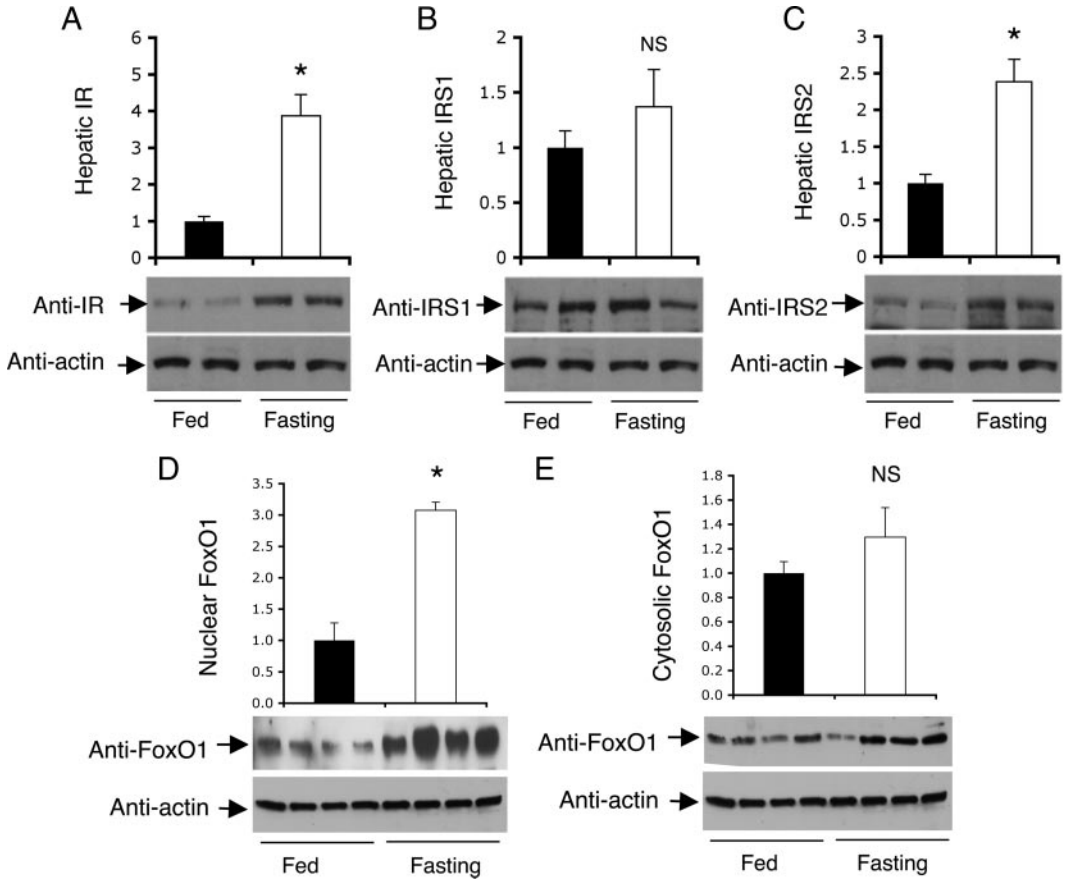


FIG. 2. Hepatic IR and IRS expression in fed and fasting states. Male CD1 mice (10 wk old) were fasted for 16 h (n = 6) or fed *ad libitum* (n = 6). Liver tissue was collected from killed mice for the preparation of liver protein extracts. Aliquots of 20-μg liver proteins were subjected to immunoblot assay. Hepatic IR (A), IRS1 (B), and IRS2 (C) protein levels were determined. Additional aliquots of liver tissue (20 mg) were homogenized for the preparation of nuclear and cytoplasmic fractions, followed by the determination of FoxO1 in the nucleus (D) and cytoplasm (E) of hepatocytes in fed and fasted mice. *, $P < 0.05$ *vs.* control. NS, Not significant.

function exhibited significantly reduced glycogen content (Fig. 3E).

Impact of FoxO1 on hepatic ER stress

To further understand the physiology underlying FoxO1-mediated feedback regulation of hepatic insulin signaling, we investigate the impact of FoxO1-ADA on hepatic ER stress, a cellular response that is associated with altered hepatic metabolism in obesity and diabetes (37, 39, 41, 42, 50, 51). As shown in Fig. 4, hepatic FoxO1-ADA production resulted in a selective induction of GRP78 and CCAAT-enhancer-binding protein homology protein (CHOP). In contrast, no significant alterations were seen in hepatic expression of other functions involved in ER stress, including the activating transcription factor (ATF)4, ER degradation enhancer, mannosidase α-like 1, growth arrest and DNA damage-inducing protein 34, inositol requiring 1 (IRE1), and protein kinase R-like ER kinase (PERK). Because GRP78 is an ER stress sensor, we chose to focus our studies on GRP78. We confirmed that GRP78 protein levels were also significantly up-regulated in response to FoxO1-ADA production in liver (Fig. 4H). These findings raised the hypothesis that FoxO1 plays a significant role in coupling hepatic insulin action to ER stress. Implicit in this hypothesis is the presence of two tandem IRE within the mouse GRP78 promoter (Fig. 5A). Likewise, four tandem IRE motifs were detected in the human GRP78 promoter (Supplemental Fig. 1), which is suggestive of an evolutionally conserved mechanism.

To address this hypothesis, we cloned the mouse GRP78 promoter in a luciferase reporter assay system in pGL3. The resulting plasmid pGRP78 was transfected to HepG2 cells in the presence or absence of FoxO1-ADA production. As shown in Fig. 5B, FoxO1-ADA production stimulated GRP78 promoter activity in a dose-dependent manner.

FoxO1 mediates insulin action on hepatic GRP78 expression

To test the ability of FoxO1 to mediate insulin action on GRP78 expression, we transfected pGRP78 into HepG2 cells that were pretransduced with adenoviral vectors expressing wild-type FoxO1 or FoxO1-ADA mutant in the presence or absence of insulin (100 nM) in culture medium. After a 24-h incubation, cells were harvested for determina-

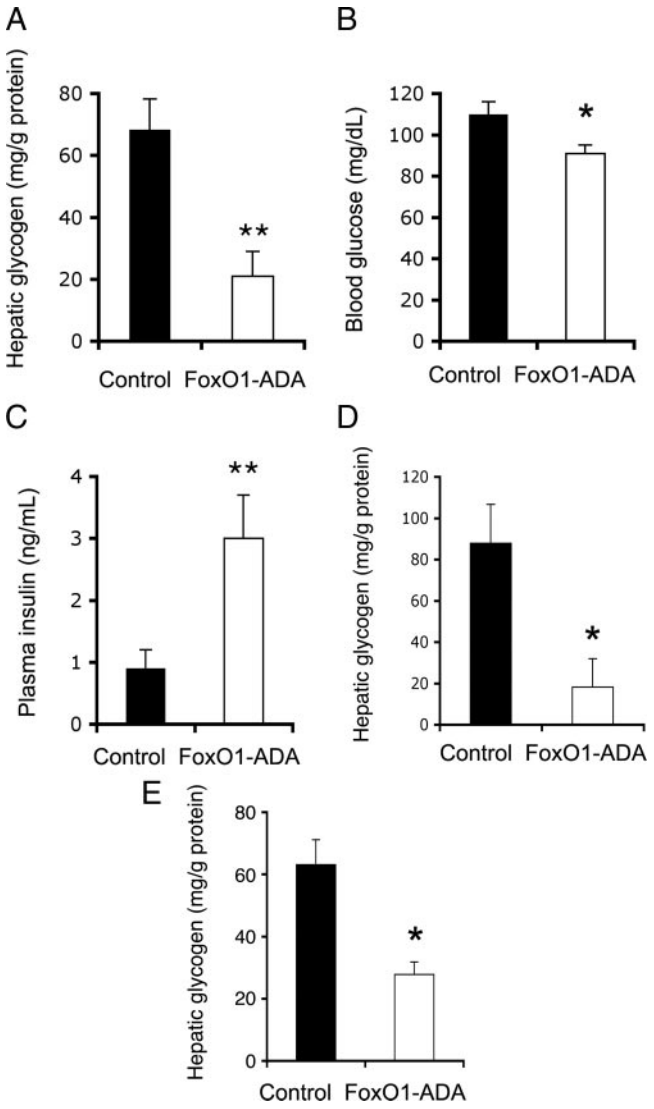


FIG. 3. Effect of FoxO1-ADA on hepatic glycogen content. Mice were killed after a 16-h fasting after 1 wk of hepatic FoxO1-ADA production as described in Fig. 1. Aliquots of liver tissues (40 mg) were used for the determination of hepatic glycogen content in FoxO1-ADA (n = 8) and control (n = 8) groups (A). Fasted blood samples were used for the determination of fasting blood glucose levels (B) and plasma insulin levels (C). Human primary hepatocytes were transduced with Adv-null or Adv-FoxO1-ADA vector (MOI, 100 pfu/cell) in 12-well microplates. Each condition was run in six replicates. After a 16-h incubation, cells were harvested for the determination of hepatic glycogen content (D). Likewise, mouse primary hepatocytes were transduced with Adv-null or Adv-FoxO1-ADA vector (MOI, 100 pfu/cell) in six-well microplates. Each condition was run in 10 replicates. After a 16-h incubation, cells were harvested for the determination of hepatic glycogen content (E). *, $P < 0.05$ and **, $P < 0.001$ *vs.* control.

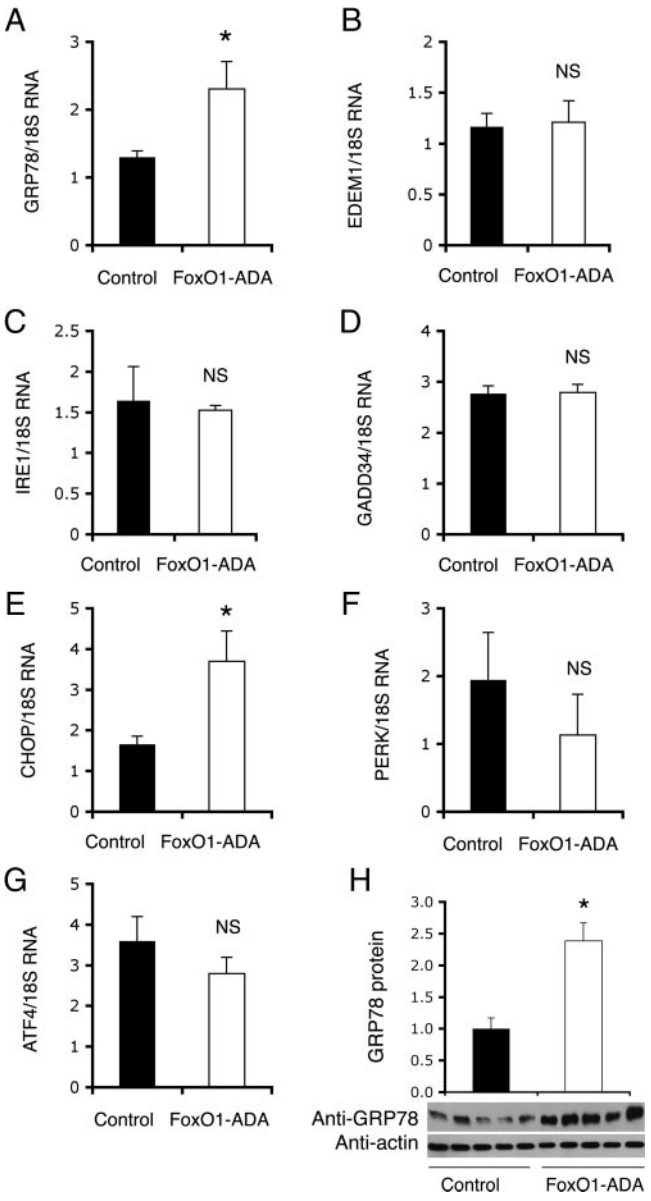


FIG. 4. Effect of FoxO1-ADA on hepatic ER stress responses. Aliquots of liver tissues (30 mg) were used for the preparation of total RNA, which was subjected to real-time quantitative RT-PCR assay for determining hepatic mRNA levels of GRP78 (A), ER degradation enhancer, mannosidase α-like (B), IRE1 (C), growth arrest and DNA damage-inducing protein 34 (D), CHOP (E), PERK (F), and ATF4 (G). Additional aliquots of liver tissue were homogenized for the preparation of total hepatic proteins, followed by immunoblot assay for the determination of hepatic GRP78 protein levels (H). *, $P < 0.01$ *vs.* control. NS, Not significant.

tion of luciferase activity. As shown in Fig. 5C, FoxO1 production resulted in a 4-fold induction of GRP78 promoter activity. This effect was counteracted by insulin, consistent with the ability of insulin to promote FoxO1 phosphorylation and translocation from the nucleus to cytoplasm (1, 2, 6). FoxO1-ADA production also contributed to a significant induction (5-fold) of GRP78 promoter activity. Unlike its wild-type counterpart, FoxO1-ADA mediated induction of GRP78 promoter activity was indifferent to insulin inhibition (Fig. 5C). This effect correlated with the inability of FoxO1-ADA to undergo insulin-dependent phosphorylation and nuclear exclusion (1, 2, 6).

FoxO1 targets GRP78 promoter for *trans*-activation

To test the hypothesis that FoxO1 targets *GRP78* gene for *trans*-activation, we mapped FoxO1 target site within the GRP78 promoter. We generated a series of promoter

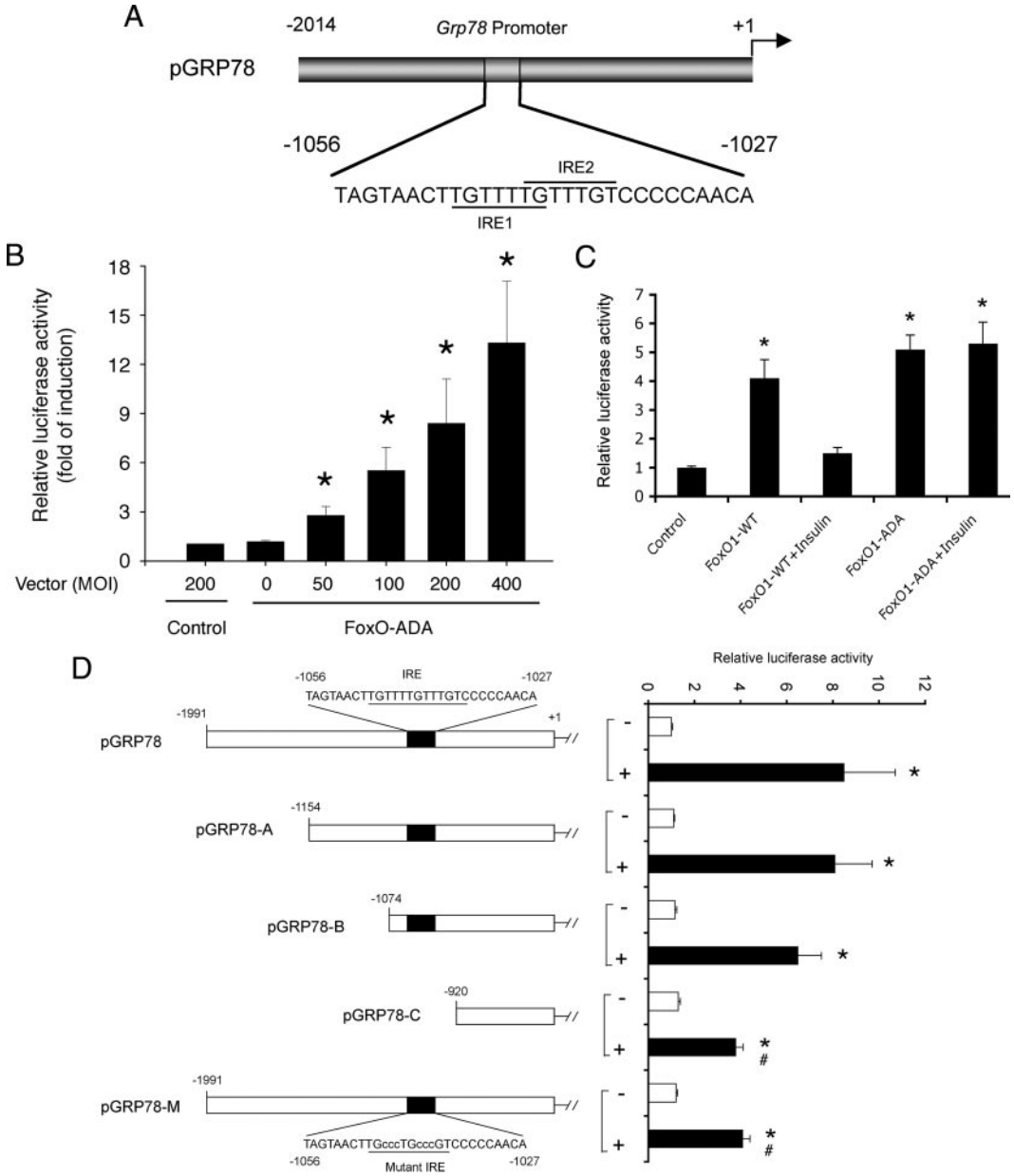


FIG. 5. Effect of FoxO1 on hepatic ER stress responses. *Schematic depiction* of the mouse GRP78 promoter. The two intertwined IRE sequences (IRE1 and IRE2) were highlighted (A). FoxO1 dose-dependent induction of GRP78 promoter activity (B). HepG2 cells were cotransfected with 2-μg pGRP78 encoding GRP78 promoter-directed luciferase reporter system and 2-μg pCMV-LacZ vector in the presence of Adv-FoxO1-ADA vector at different doses ranging from 50 to 400 pfu/cell or Adv-null at 200 pfu/cell in culture medium. After a 24-h incubation, cells were assayed for luciferase and β-galactosidase (β-gal) activities. The relative promoter activity, defined as the ratio of luciferase to β-gal activities, was determined (B). Effect of insulin on GRP78 promoter activity. HepG2 cells were cotransfected with 2-μg pGRP78 encoding GRP78 promoter-directed luciferase reporter system and 2-μg pCMV-LacZ vector in the presence of wild-type FoxO1 (FoxO1-WT) or its constitutive active allele FoxO1-ADA production. After a 24-h incubation in the presence or absence of insulin (100 nM), cells were assayed for determining the relative luciferase activity using β-gal activity as a control (C). Characterization of FoxO1 target site in GRP78 promoter. HepG2 cells were cotransfected with pCMV-LacZ plus either wild-type or mutant GRP78 promoter-directed luciferase reporter system in the presence of Adv-null or Adv-FoxO1-ADA vector at a fixed dose (100 pfu/cell). The relative luciferase activity for each construct was determined after a 24-h incubation (D). Data were obtained from five to eight experiments. *, $P < 0.001$ *vs.* basal states; #, $P < 0.05$ *vs.* control wild-type promoter construct.

variants with deletions of the upstream region of the GRP78 promoter. Using the luciferase reporter assay, we determined the activity of promoter variants in HepG2 cells in the presence and absence of FoxO1-ADA production. As shown in Fig. 5D, deletion of DNA up to −1074 nucleotide (nt) into the upstream region of GRP78 promoter did not affect promoter activity after FoxO1-ADA production in HepG2 cells. Further deletion up to −920 nt in the GRP78 promoter resulted in a significant reduction in promoter activity in response to FoxO1-ADA production. These results are in line with the presence of two IRE motifs conjoined within the region (−1074/−920 nt) of

the GRP78 promoter. To strengthen these findings, we altered the IRE sequence in the GRP78 promoter by site-directed mutagenesis. After sequencing confirmation (Supplemental Fig. 2), the resulting mutant promoter was assessed for its ability to respond to FoxO1-ADA production. As expected, mutations in the IRE region significantly attenuated the promoter activity in response to FoxO1-ADA induction (Fig. 5D).

Molecular association of FoxO1 with GRP78 promoter DNA

To consolidate the above results, we performed EMSA to visualize the molecular association between FoxO1 and GRP78 IRE DNA. We prepared FoxO1-enriched nuclear protein extract from FoxO1-expressing HepG2 cells as described (25). Aliquots of FoxO1-containing protein extract were incubated with a 24-bp GRP78 IRE DNA sequence that was prelabeled with biotin, followed by chemiluminescent EMSA. As shown in Fig. 6A, the migration of the GRP78 IRE DNA was significantly retarded in the presence of FoxO1 in 6% native polyacrylamide gels. Addition of anti-FoxO1 antibody to the reaction mixture resulted in a supershifted DNA band. To confirm the specificity of FoxO1-IRE DNA interaction, we included nonlabeled IRE DNA at 100-fold higher concentrations as competitors in the reaction, demonstrating that IRE DNA shift was abolished. As control, we performed EMSA with a mutant IRE DNA containing six base substitutions. No shifted and supershifted DNA bands were detectable in the EMSA, indicating that mutations in the GRP78 IRE motif abrogated its ability to associate with FoxO1 protein.

To further underpin these data, we employed chromatin immunoprecipitation (ChIP) assay to examine the molecular interaction between FoxO1 and the GRP78 promoter. We transfected pGRP78 into HepG2 cells that were pretransduced with FoxO1 vector, followed by ChIP assay using rabbit anti-FoxO1 antibody or preimmune rabbit sera. The immunoprecipitates were subjected to immunoblot assay for detecting immunoprecipitated FoxO1 and PCR analysis for visualizing coimmunoprecipitated DNA. As shown in Fig. 6B, specific DNA corresponding to the proximal region (−1074/+1 nt) of the GRP78 promoter was amplified by PCR using primers flanking the GRP78 IRE motif. In contrast, the immunoprecipitates derived from preimmune sera were negative in the same PCR assay. As an input control, aliquots of cell lysates (1 μl) before immunoprecipitation were analyzed. Specific DNA bands corresponding to the GRP78 promoter were detected (Fig. 6B). In addition, we performed PCR analysis using a pair of off-target primers flanking a distal region (−4671/−4652 nt) that is devoid of the consensus IRE motif in the GRP78 promoter. No specific DNA was amplified in the immunoprecipitates derived from preimmune IgG or anti-FoxO1 antibody (Fig. 6B).

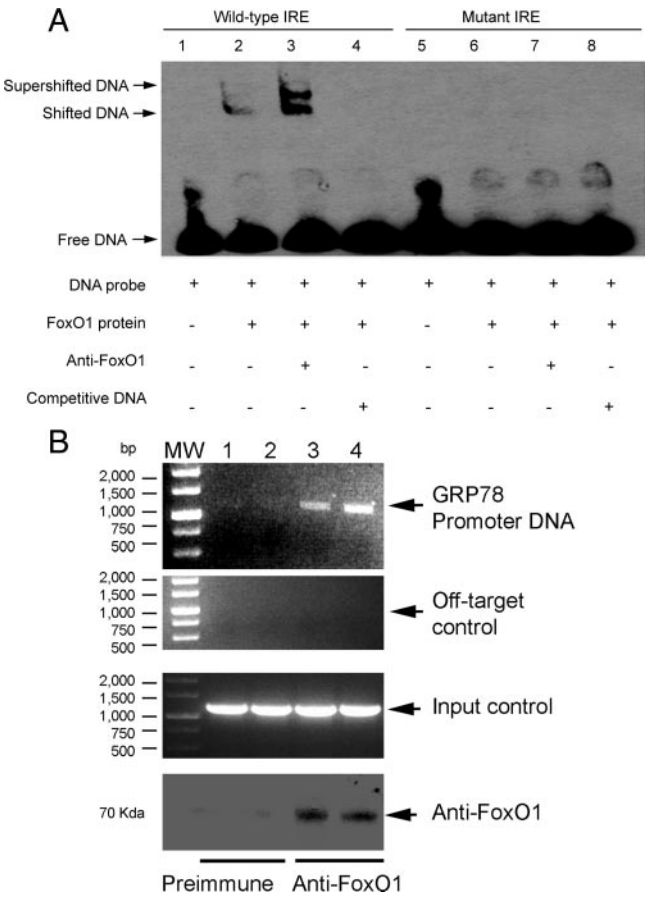


FIG. 6. Molecular association of FoxO1 with GRP78 promoter. FoxO1 binding to DNA. Aliquots of FoxO1 protein extracts (5 μg) were incubated with biotin-labeled DNA probe, followed by chemiluminescent EMSA. FoxO1 protein lysates were prepared from HepG2 cells that were pretransduced with FoxO1 vector (MOI, 200 pfu/cell). DNA probe was derived from a 24-bp DNA covering the consensus IRE (−1056/−1023 nt) of the mouse GRP78 promoter (lanes 1–4). A mutant DNA probe with altered IRE motif was used as a control (lanes 5–8) in the EMSA (A). ChIP assay. HepG2 cells were transfected with pGRP78 in the presence of FoxO1 vector at an MOI of 100 pfu/cell in triplicate. After a 24-h incubation, cells were cross-linked with 1% formaldehyde, followed by ChIP assay using preimmune rabbit IgG (lane 1 and 2) or rabbit anti-FoxO1 antibody (lanes 3 and 4). Immunoprecipitates were subjected to PCR analysis using a pair of primers flanking the IRE sequence in the GRP78 promoter. As a negative control, the immunoprecipitates were subjected to PCR analysis using a pair of off-target primers flanking a distal region (−4671/−4652 nt) that is devoid of the consensus IRE motif in the upstream region of the GRP78 promoter. As a positive control, aliquots of input DNA samples (1 μl) were used in PCR analysis. In addition, aliquots of immunoprecipitates were subjected to anti-FoxO1 immunoblot analysis for confirming the presence of FoxO1 protein. Data were from three independent repeats (B). MW, molecular weight.

FoxO1 interacts with GRP78 promoter in liver

To recapitulate the above finding *in vivo*, we performed ChIP assay on liver tissues of fed and fasted mice. Because FoxO1 activity is induced in fasted liver, we reasoned that

this effect would translate into an induction of GRP78 expression in fasted mice. As shown in Fig. 7, A–C, positive association of FoxO1 with GRP78 promoter DNA was detected in liver. This effect was induced in mice after a 16-h fast, correlating with a significant induction of hepatic GRP78 mRNA levels in fasted mice.

To ascertain the finding of increased FoxO1 activity in fasted liver, we subjected liver tissue of fed and fasted mice to immunohistochemistry. FoxO1 was predominantly localized in the nucleus of hepatocytes in fasted mice (Fig. 7, D–F). In contrast, FoxO1 was expressed at basal level under fed conditions (Fig. 7, G–I).

To reinforce the idea that FoxO1 targets *GRP78* gene for *trans*-activation, we investigated the interaction of FoxO1 with GRP78 promoter in insulin resistant liver of obese *db/db* mice. When compared with heterozygous lean *db/+* mice, obese *db/db* mice exhibited significantly increased GRP78 expression (Fig. 7J), accompanied by a 5-fold induction of FoxO1 expression in liver (Fig. 7K). This effect correlated with a marked induction in molecular association between FoxO1 and GRP78 promoter DNA in insulin resistant liver of obese *db/db* mice (Fig. 7, L and M). These results are consistent with our previous observations that FoxO1 becomes deregulated in insulin resistant liver, as reflected in its increased nuclear redistribution in hepatocytes of *db/db* mice (24, 25).

FoxO1 links saturated fat, but not thapsigargin, to ER stress

To further illustrate the underlying pathophysiology of FoxO1-mediated induction of GRP78 expression, we determined the expression levels of GRP78 and FoxO1 proteins in HepG2 cells that pretreated with palmitate (250 μM), the predominant saturated form of FFA that is deleterious to hepatic insulin signaling (52–55). Palmitate treatment resulted in a significant induction of GRP78 protein expression, which is indicative of ER stress in HepG2 cells (Fig. 8A). This effect correlated with a 2-fold induction of FoxO1 production in the nucleus of palmitate-treated HepG2 cells (Fig. 8, B and C). In accordance with these findings, Wei *et al.* (54) show that palmitate promotes CHOP production and elicits ER stress in cultured H4IIE cells.

To correlate FoxO1 activity with GRP78 induction, we employed small interfering RNA (siRNA)-mediated gene silencing approach to ablate FoxO1 expression in HepG2 cells, using Adv-FoxO1-siRNA vector as described (25). In response to siRNA-mediated FoxO1 knockdown, palmitate-mediated induction of GRP78 expression was significantly attenuated (Fig. 8, D and E), underscoring the importance of FoxO1 in palmitate-mediated induction of ER stress.

To further underpin the importance of FoxO1 in ER stress, we incubated palmitate-treated HepG2 cells in the absence and presence of 1-mM PBA, a pharmacological chaperone that is shown to reduce cellular ER stress and improve insulin sensitivity in rodent models of type 2 diabetes (38). PBA treatment ameliorated palmitate-elicited ER stress, as evidenced by the significant reduction of GRP78 expression (Fig. 8F). This effect was accompanied by a concomitant reduction of FoxO1 expression in palmitate-treated HepG2 cells (Fig. 8G). PBA treatment also attenuated palmitate-mediated induction of CHOP, X-box binding protein 1, ATF4, eukaryotic translation initiation factor 2 α, PERK, and ATF6 expression to different extents in HepG2 cells (Supplemental Fig. 3).

As control, we incubated HepG2 cells in the absence and presence of monounsaturated fat oleate (500 μM), followed by analysis of GRP78 and FoxO1 expression. In contrast to palmitate, oleate treatment resulted in a relatively milder induction of GRP78 expression (~30%) (Fig. 8H), accompanied by a small increase of nuclear FoxO1 protein levels in oleate-treated HepG2 cells (Fig. 8, I and J).

Thapsigargin is a potent inducer of ER stress. It raises cytosolic calcium concentration by blocking the ability of cells to pump calcium into ER lumen, resulting in overt ER stress (56). To test whether FoxO1 is responsible for thapsigargin-induced ER stress, we cultured HepG2 cells in the absence and presence of thapsigargin (250 nM), followed by determination of FoxO1 and GRP78 expression levels. As expected, thapsigargin treatment resulted in about a 6-fold induction of GRP78 expression (Fig. 8K). In contrast, no significant differences in FoxO1 expression levels were detected in control and thapsigargin-treated HepG2 cells (Fig. 8, L and M). Together, these data suggest that FoxO1 played a significant role in coupling saturated fat, but not thapsigargin, to ER stress.

Discussion

FoxO1 has emerged as an important transcriptional factor that integrates hepatic insulin signaling to target genes in hepatic metabolism. Abundantly expressed in liver, FoxO1 controls insulin-dependent inhibition of PEPCK and G6PC, two key enzymes that are involved in gluconeogenesis (1, 2). In response to fasting, FoxO1 activity is enhanced, culminating in its increased nuclear localization in liver. This effect stimulates PEPCK and G6PC expression and promotes hepatic glucose production to maintain fasting blood sugar levels within the physiological range (1, 2). In response to refeeding, FoxO1 undergoes insulin-dependent phosphorylation and nuclear exclusion, resulting in inhibition of PEPCK and G6PC production in liver.

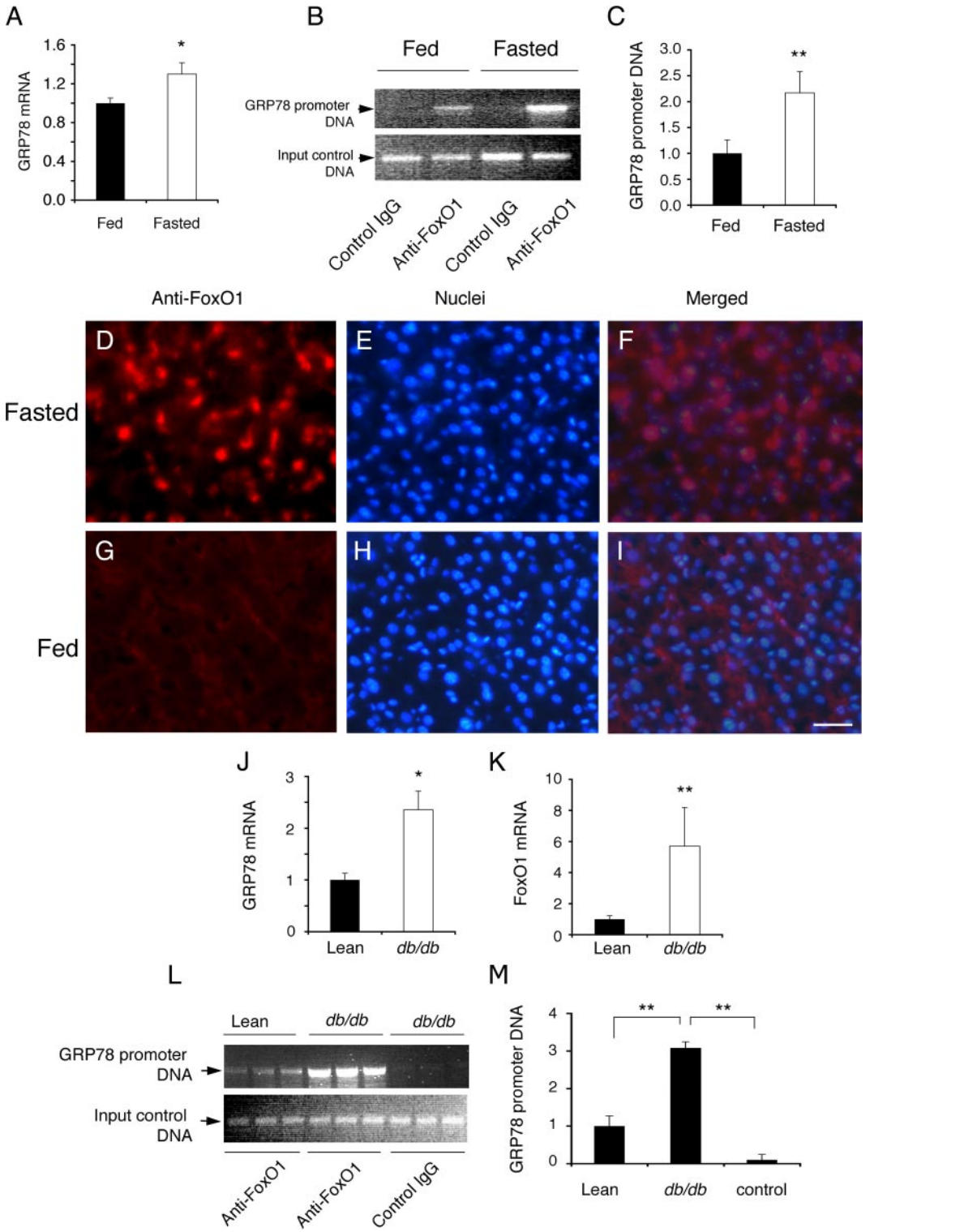


FIG. 7. FoxO1 association with GRP78 in liver. Male CD1 mice (10 wk old, n = 6 per group) were fasted for 16 h or fed *ad libitum*. Liver tissue was collected from killed mice for the preparation of total hepatic RNA, followed by real-time quantitative RT-PCR assay for the determination of GRP78 mRNA levels (A). Aliquots of liver tissue (20 mg) were subjected to ChIP assay using anti-FoxO1 or control anti-β-galactosidase antibodies. Immunoprecipitates were analyzed by PCR using specific primers flanking the IRE DNA motif in the GRP78 promoter in fed and fasted liver (B). The relative amount of GRP78 promoter DNA immunoprecipitated from fed and fasted liver was determined (C). Additionally, frozen liver sections (8 μm) were immunostained with anti-FoxO1 antibody for visualizing FoxO1 in the liver under fasting (D–F) and fed (G–I) conditions. Male obese *db/db* (16 wk old, n = 6 per group) and age-/sex-matched lean *db/+* littermates were killed, and liver tissue (20 mg) was used for the preparation of total hepatic RNA, followed by real-time quantitative RT-PCR analysis for determining GRP78 (J) and FoxO1 (K) expression. Furthermore, aliquots of liver tissue (20 mg) were subjected to ChIP assay using anti-FoxO1 or control anti-β-galactosidase antibodies. Immunoprecipitates were analyzed by PCR using specific primers flanking the IRE DNA motif in the GRP78 promoter (L). The relative amount of GRP78 promoter DNA immunoprecipitated from the liver of lean and obese *db/db* mice was determined (M). *, $P < 0.05$; **, $P < 0.001$ *vs.* control. *Scale bar* in D–I, 20 μm.

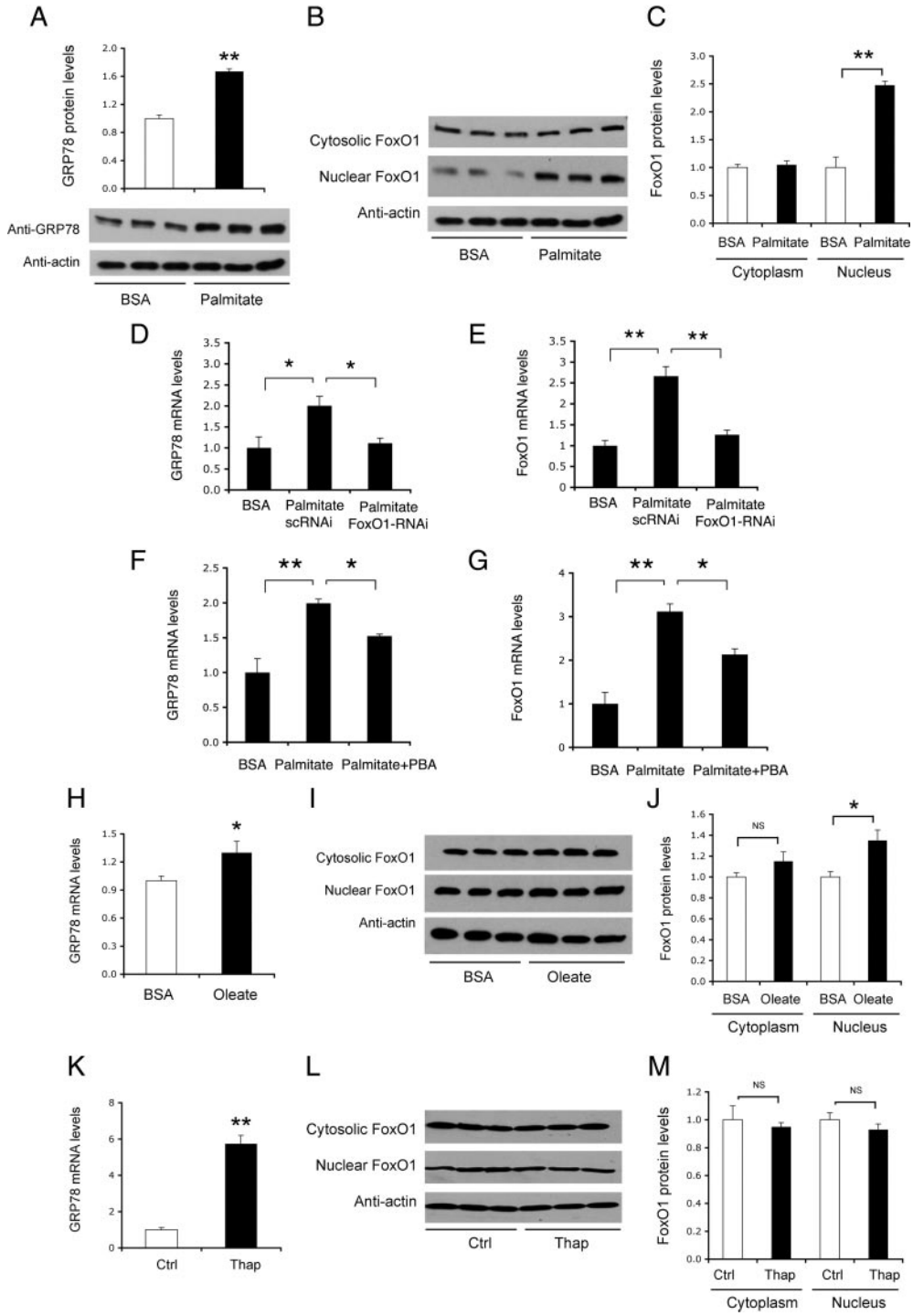


FIG. 8. FoxO1 links palmitate, but not thapsigargin, to ER stress. HepG2 cells were treated with 250 μM of BSA-bound palmitate or BSA for 16 h (n = 3 for each condition), followed by immunoblot analysis using anti-GRP78 antibody (A). In addition, treated cells were subjected to nuclear and cytosolic fractionation. Aliquots (20 μg) of nuclear and cytosolic proteins were analyzed by anti-FoxO1 immunoblot assay (B), followed by the determination of cytosolic *vs.* nuclear FoxO1 protein levels (C). Separately, palmitate-treated cells were transduced with 100 MOI of Adv-FoxO1-RNAi encoding FoxO1-specific siRNA or control Adv-FoxO1-scRNAi vector encoding scrambled siRNA. After a 16-h incubation, cells were subjected to real-time quantitative (q)RT-PCR analysis for determining GRP78 (D) and FoxO1 (E) expression. Similarly, HepG2 cells were treated with palmitate (250 μM) in the presence and absence of PBA (1 mM) in culture medium. After a 2-h incubation, cells were subjected to real-time qRT-PCR analysis for the determination of GRP78 (F) and FoxO1 (G) expression. Likewise, HepG2 cells were treated with 500 μM of BSA-bound oleate or BSA for 16 h (n = 3 for each condition), followed by real-time qRT-PCR assay using specific primers of GRP78 or 18S rRNA (H). Treated cells were subjected to anti-FoxO1 immunoblot assay (I) for the determination of cytosolic *vs.* nuclear FoxO1 protein levels (J). Additionally, HepG2 cells were treated with 250 nM of thapsigargin for 16 h. Control cells were mock treated (n = 3 for each condition). Treated cells were subjected to real-time qRT-PCR assay for the determination of GRP78 mRNA levels (K) and to anti-FoxO1 immunoblot assay (L) for the determination of cytosolic *vs.* nuclear FoxO1 protein levels (M). *, $P < 0.05$ and **, $P < 0.001$ *vs.* control. NS, Not significant; Ctrl, control group; Thap, thapsigargin group.

This effect limits hepatic glucose production and prevents prolonged postprandial blood glucose excursion after meals. Such a reciprocal mechanism is essential for liver to adjust the rate of hepatic glucose in response to insulin and nutrient availability.

A substrate of serine-threonine kinase/protein kinase B, FoxO1 is shown to stimulate the expression of its upstream effectors IR and IRS2, setting a feedback loop that negatively regulates FoxO1 activity in liver (Supplemental Fig. 4). It has been proposed that such a feedback loop serves as an adaptive mechanism to enhance insulin sensitivity in fasting states to program starved cells for forthcoming nutrients (29). This notion seems counterintuitive, because fasting tends to desensitize peripheral tissues to insulin to minimize carbohydrate metabolism in favor of survival in the face of famine (31). In this study, we probed the biological consequence of a dislodged FoxO1 feedback loop for better understanding of the physiology that underlies FoxO1-mediated feedback regulation of IR and IRS2 in liver. Using adenovirus-mediated gene transfer approach, we achieved hepatic production of FoxO1-ADA, a constitutive active allele that is able to evade IR- and IRS2-facilitated feedback regulation. We show that hepatic FoxO1-ADA production 1) resulted in a significant induction of IR and IRS expression, 2) caused near depletion of hepatic glycogen content, and 3) induced GRP78 production in liver. These findings were recapitulated in cultured HepG2 cells, as well as human and mouse primary hepatocytes with elevated FoxO1 activity. Furthermore, we show that hepatic IR and IRS2 expression was significantly up-regulated, correlating with the induction of FoxO1 activity in liver of fasted mice. Our data corroborate the idea that hepatic FoxO1 activity is subject to feedback regulation in an IR- and IRS2-dependent manner. We illustrate that the FoxO1 feedback loop plays an important role in limiting hepatic FoxO1 activity to prevent potential glycogen depletion and ER stress in liver (Supplemental Fig. 4).

Another important finding derived from this study is the revelation of the mechanism by which FoxO1 mediates insulin-dependent regulation of GRP78, an ER stress sensor molecule. FoxO1 targets *GRP78* gene for *trans*-activation, and this effect is counteracted by insulin. Although ER stress is closely associated with insulin resistance, the underlying mechanism remains obscure. We show that FoxO1 activity is increased, accompanied by a significant induction of GRP78 expression in insulin resistant liver of obese *db/db* mice. Our interpretation is that in response to insulin resistance, hepatic FoxO1 activity is enhanced, culminating in its increased nuclear localization. This effect acts to stimulate hepatic production of GRP78, which in turn functions to resolve ER stress in liver. Thus, FoxO1-mediated induction of GRP78 plays an important role in adaptive UPR activation in response to attenuated insulin action in liver. Consistent with this notion is the observation that hepatic FoxO1 activity is significantly elevated (17), correlating with a marked induction of GRP78 expression in liver of high fat-induced obese mice (50). Weight loss-mediated improvement in insulin sensitivity is associated with the reduction of hepatic GRP78 expression and ER stress in obese subjects (41). Adenovirus-mediated hepatic GRP78 overproduction is sufficient to mitigate ER stress and enhance hepatic insulin sensitivity in diabetic *db/db* mice (37).

However, this view is challenged by a recent study showing that GRP78 haploinsufficiency attenuates high fat-induced insulin resistance and obesity in C57BL/6J mice, implicating a direct role of GRP78 in ER stress (57). It is noteworthy that GRP78 haploinsufficiency also promotes chronic UPR, resulting in a compensatory induction of other chaperones, such as GRP94 and protein disulfide isomerase in GRP78+/− heterozygous mice. This compensatory mechanism along with augmented residual UPR caused by GRP78 loss-of-function may contribute to the amelioration of diet-induced obesity and insulin resistance in GRP78+/− heterozygous mice.

Although thapsigargin, palmitate, and oleate elicited variable degrees of ER stress, as reflected by increased GRP78 expression in HepG2 cells, only palmitate-mediated ER stress was coupled with a marked induction of FoxO1 production. These results are novel, suggesting that FoxO1-mediated induction of GRP78 production and UPR is specific to overload of polyunsaturated fat. These results are in line with the observation by Kamagate *et al.* (25), who report that in response to increased lipid load, FoxO1 activity is enhanced, which in turn promotes hepatic MTP production and VLDL-triglyceride secretion. This raises a fundamental question: Why the liver cannot rid itself of excessive lipids, and avoid hepatic ER stress and steatosis by accelerating VLDL secretion in the face of lipid excess in subjects with visceral obesity or type 2 diabetes? An important insight to this question is gained from the study by Ota *et al.* (39), who show that hepatic VLDL production is sensitive to ER stress in a parabolic manner. Moderate ER stress does induce VLDL-triglyceride secretion, protecting liver from ER stress-induced steatosis. However, excessive ER stress in response to prolonged exposure to lipids impairs the ability of liver to secrete triglycerides. This effect contributes to lipid accumulation in liver, exacerbating hepatic steatosis (39). Indeed, this lipid-induced hepatic ER stress is concomitant with steatosis in both genetic and dietary models of obese mice (39, 50, 58), as well as in high fructose-fed hamsters (33, 59, 60).

Although FoxO1 was shown to bind and stimulate GRP78 promoter activity, neither deletion nor mutations

of the consensus FoxO1 binding site abolished GRP78 promoter activity. These results are suggestive of additional mechanisms that may account for increased GRP78 production and ER stress in response to increased FoxO1 activity in HepG2 cells. Further studies are warranted to dissect the underlying mechanism of FoxO1-mediated regulation of GRP78 expression for better understanding of the molecular basis that couples ER stress to insulin resistance in obesity and type 2 diabetes.

An ER-resident protein, GRP78 remains bound to IRE1, PERK, and ATF6 in unstressed cells. In response to the accumulation of misfolded proteins in ER lumen, GRP78 dissociates from IRE1, PERK, and ATF6, triggering UPR for attenuating the rate of protein synthesis and promoting the induction of genes encoding ER chaperones (35, 49, 61). Thus, GRP78 is hailed as an ER chaperone for sensing stress signal and mounting UPR to resolve ER stress. GRP78 also plays a critical role in targeting misfolded proteins for proteasomal degradation, which is reviewed as an ER quality-control mechanism (40, 61, 62). In addition to its sensitivity to disruption in protein folding, the ER lumen is vulnerable to alterations in oxidizing redox potential (63), luminal calcium homeostasis (64), and excessive lipid accumulation (39, 65). Thus, there are multiple routes leading to the induction of ER stress. Unresolved ER stress is deleterious to cell growth and metabolism (41, 66–68). Mice with genetic ablation of the regulatory subunit p85α of phosphatidylinositol kinase in the liver exhibit impaired hepatic insulin action, accompanied by profound ER stress in response to tunicamycin administration (48). Chemical chaperone-mediated inhibition of ER stress is shown to improve glucose metabolism and enhance insulin sensitivity in type 2 diabetic mice (38).

In conclusion, our data consolidate the idea that hepatic FoxO1 activity is subject to feedback regulation. Unchecked FoxO1 activity, resulting from molecular defects in the FoxO1 feedback loop, is deleterious to hepatic metabolism, culminating in unrestrained glycogen breakdown and excessive ER stress in liver. Previous studies show that FoxO1 plays a pivotal role in mediating insulin-dependent regulation of hepatic glucose and VLDL production. FoxO1 dysregulation, resulting from an impaired ability to curb FoxO1 activity, is attributable to hepatic glucose and triglyceride overproduction, accounting in part for the dual pathogenesis of fasting hyperglycemia and hypertriglyceridemia in insulin resistant subjects with obesity and/or type 2 diabetes (1, 2, 5, 6, 27, 34). Our present data, together with previous findings, suggest that the FoxO1 feedback loop may serve as a safeguarding mechanism for keeping FoxO1 activity in check to avert hepatic glycogen depletion and ER stress.

Acknowledgments

We thank Dr. Domenico Accili (Columbia University, New York, NY) for providing FoxO1 vectors.

Address all correspondence and requests for reprints to: H. Henry Dong, Rangos Research Center, Children's Hospital of Pittsburgh of University of Pittsburgh Medical Center, 4401 Penn Avenue, Pittsburgh, Pennsylvania 15224. E-mail: dongh@pitt.edu.

This work was supported in part by American Diabetes Association and National Institutes of Health (NIH) Grant DK066301. S.C.S. and R.G. were supported by the NIH Contract N01-DK-7-0004/HHSN26700700004C for the Liver Tissue Cell Distribution System at the University of Pittsburgh School of Medicine.

Disclosure Summary: The authors have nothing to disclose.

References

1. **Accili D, Arden KC** 2004 FoxOs at the crossroads of cellular metabolism, differentiation, and transformation. Cell 117:421–426
2. **Barthel A, Schmoll D, Unterman TG** 2005 FoxO proteins in insulin action and metabolism. Trends Endocrin Met 16:183–189
3. **Lee SS, Kennedy S, Tolonen AC, Ruvkun G** 2003 DAF-16 target genes that control *C. elegans* life-span and metabolism. Science 300: 644–647
4. **Hwangbo DS, Gershman B, Tu MP, Palmer M, Tatar M** 2004 Drosophila dFOXO controls lifespan and regulates insulin signalling in brain and fat body. Nature 429:562–566
5. **Dong XC, Copps KD, Guo S, Li Y, Kollipara R, DePinho RA, White MF** 2008 Inactivation of hepatic Foxo1 by insulin signaling is required for adaptive nutrient homeostasis and endocrine growth regulation. Cell Metab 8:65–76
6. **Kamagate A, Dong HH** 2008 FoxO1 integrates insulin signaling to VLDL production. Cell Cycle 7:3162–3170
7. **Biggs 3rd WH, Meisenhelder J, Hunter T, Cavenee WK, Arden KC** 1999 Protein kinase B/Akt-mediated phosphorylation promotes nuclear exclusion of the winged helix transcription factor FKHR1. Proc Natl Acad Sci USA 96:7421–7426
8. **Rena G, Prescott AR, Guo S, Cohen P, Unterman TG** 2001 Roles of the forkhead in rhabdomyosarcoma (FKHR) phosphorylation sites in regulating 14-3-3 binding, transactivation and nuclear targetting. Biochem J 354:605–612
9. **Nakae J, Park BC, Accili D** 1999 Insulin stimulates phosphorylation of the forkhead transcription factor FKHR on serine 253 through a wortmannin-sensitive pathway. J Biol Chem 274:15982–15985
10. **Nakae J, Barr V, Accili D** 2000 Differential regulation of gene expression by insulin and IGF-1 receptors correlates with phosphorylation of a single amino acid residue in the forkhead transcription factor FKHR. EMBO J 19:989–996
11. **Durham SK, Suwanichkul A, Scheimann AO, Yee D, Jackson JG, Barr FG, Powell DR** 1999 FKHR binds the insulin response element in the insulin-like growth factor binding protein-1 promoter. Endocrinology 140:3140–3146
12. **Schmoll D, Walker KS, Alessi DR, Grempler R, Burchell A, Guo S, Walther R, Unterman TG** 2000 Regulation of glucose-6-phosphatase gene expression by protein kinase B α and the forkhead transcription factor FKHR. J Biol Chem 36324–36333
13. **Zhang X, Gan L, Pan H, Guo S, He X, Olson ST, Mesecar A, Adam S, Unterman TG** 2002 Phosphorylation of serine 256 suppresses transactivation by FKHR (FOXO1) by multiple mechanisms. Direct and indirect effects on nuclear/cytoplasmic shuttling and DNA binding. J Biol Chem 277:45276–45284

14. **Van Der Heide LP, Hoekman MF, Smidt MP** 2004 The ins and outs of FoxO shuttling: mechanisms of FoxO translocation and transcriptional regulation. Biochem J 380:297–309
15. **van der Heide LP, Jacobs FM, Burbach JP, Hoekman MF, Smidt MP** 2005 FoxO6 transcriptional activity is regulated by Thr26 and Ser184, independent of nucleo-cytoplasmic shuttling. Biochem J 391:623–629
16. **Rena G, Guo S, Cichy SC, Unterman TG, Cohen P** 1999 Phosphorylation of the transcription factor forkhead family member FKHR by protein kinase B. J Biol Chem 274:17179–17183
17. **Qu S, Altomonte J, Perdomo G, He J, Fan Y, Kamagate A, Meseck M, Dong HH** 2006 Aberrant forkhead box O1 function is associated with impaired hepatic metabolism. Endocrinology 147:5641–5652
18. **Zhao X, Gan L, Pan H, Kan D, Majeski M, Adam SA, Unterman TG** 2004 Multiple elements regulate nuclear/cytoplasmic shuttling of FOXO1: characterization of phosphorylation- and 14-3-3-dependent and -independent mechanisms. Biochem J 378:839–849
19. **Nakae J, Kitamura T, Silver DL, Accili D** 2001 The forkhead transcription factor Foxo1 (Fkhr) confers insulin sensitivity onto glucose-6-phosphatase expression. J Clin Invest 108:1359–1367
20. **Altomonte J, Richter A, Harbaran S, Suriawinata J, Nakae J, Thung SN, Meseck M, Accili D, Dong H** 2003 Inhibition of Foxo1 function is associated with improved fasting glycemia in diabetic mice. Am J Physiol 285:E718–E728
21. **Zhang W, Patil S, Chauhan B, Guo S, Powell DR, Le J, Klotsas A, Matika R, Xiao X, Franks R, Heidenreich KA, Sajan MP, Farese RV, Stolz DB, Tso P, Koo SH, Montminy M, Unterman TG** 2006 FoxO1 regulates multiple metabolic pathways in the liver: effects on gluconeogenic, glycolytic, and lipogenic gene expression. J Biol Chem 281:10105–10117
22. **Puigserver P, Rhee J, Donovan J, Walkey CJ, Yoon JC, Oriente F, Kitamura Y, Altomonte J, Dong H, Accili D, Spiegelman BM** 2003 Insulin-regulated hepatic gluconeogenesis through FOXO1-PGC-1α interaction. Nature 423:550–555
23. **Matsumoto M, Pocai A, Rossetti L, Depinho RA, Accili D** 2007 Impaired regulation of hepatic glucose production in mice lacking the forkhead transcription factor Foxo1 in liver. Cell Metab 6:208–216
24. **Altomonte J, Cong L, Harbaran S, Richter A, Xu J, Meseck M, Dong HH** 2004 Foxo1 mediates insulin action on ApoC-III and triglyceride metabolism. J Clin Invest 114:1493–1503
25. **Kamagate A, Qu S, Perdomo G, Su D, Kim DH, Slusher S, Meseck M, Dong HH** 2008 FoxO1 mediates insulin-dependent regulation of hepatic VLDL production in mice. J Clin Invest 118:2347–2364
26. **Matsumoto M, Han S, Kitamura T, Accili D** 2006 Dual role of transcription factor FoxO1 in controlling hepatic insulin sensitivity and lipid metabolism. J Clin Invest 116:2464–2472
27. **Sparks DJ, Dong HH** 2009 FoxO1 and hepatic lipid metabolism. Curr Opin Lipidol 20:217–226
28. **Sparks JD, Sparks CE** 2008 Overindulgence and metabolic syndrome: is FoxO1 a missing link? J Clin Invest 118:2012–2015
29. **Puig O, Tjian R** 2005 Transcriptional feedback control of insulin receptor by dFOXO/FOXO1. Gene Dev 19:2435–2446
30. **Puig O, Tjian R** 2006 Nutrient availability and growth: regulation of insulin signaling by dFOXO/FOXO1. Cell Cycle 5:503–505
31. **van der Crabben SN, Allick G, Ackermans MT, Endert E, Romijn JA, Sauerwein HP** 2008 Prolonged fasting induces peripheral insulin resistance, which is not ameliorated by high-dose salicylate. J Clin Endocrinol Metab 93:638–641
32. **Heijboer AC, Donga E, Voshol PJ, Dang ZC, Havekes LM, Romijn JA, Corssmit EP** 2005 Sixteen hours of fasting differentially affects hepatic and muscle insulin sensitivity in mice. J Lipid Res 46:582–588
33. **Qu S, Su D, Altomonte J, Kamagate A, He J, Perdomo G, Tse T, Jiang Y, Dong HH** 2007 PPARα mediates the hypolipidemic action of fibrates by antagonizing FoxO1. Am J Physiol 292:E421–E434
34. **Biddinger SB, Hernandez-Ono A, Rask-Madsen C, Haas JT, Alemán JO, Suzuki R, Scapa EF, Agarwal C, Carey MC, Stephanopoulos G, Cohen DE, King GL, Ginsberg HN, Kahn CR** 2008 Hepatic insulin resistance is sufficient to produce dyslipidemia and susceptibility to atherosclerosis. Cell Metab 7:125–134
35. **Hotamisligil GS** 2010 Endoplasmic reticulum stress and the inflammatory basis of metabolic disease. Cell 140:900–917
36. **Bánhegyi G, Baumeister P, Benedetti A, Dong D, Fu Y, Lee AS, Li J, Mao C, Margittai E, Ni M, Paschen W, Piccirella S, Senesi S, Sitia R, Wang M, Yang W** 2007 Endoplasmic reticulum stress. Ann NY Acad Sci 1113:58–71
37. **Kammoun HL, Chabanon H, Hainault I, Luquet S, Magnan C, Koike T, Ferré P, Foufelle F** 2009 GRP78 expression inhibits insulin and ER stress-induced SREBP-1c activation and reduces hepatic steatosis in mice. J Clin Invest 119:1201–1215
38. **Ozcan U, Yilmaz E, Ozcan L, Furuhashi M, Vaillancourt E, Smith RO, Görgün CZ, Hotamisligil GS** 2006 Chemical chaperones reduce ER stress and restore glucose homeostasis in a mouse model of type 2 diabetes. Science 313:1137–1140
39. **Ota T, Gayet C, Ginsberg HN** 2008 Inhibition of apolipoprotein B100 secretion by lipid-induced hepatic endoplasmic reticulum stress in rodents. J Clin Invest 118:316–332
40. **Qiu W, Kohen-Avramoglu R, Mhapsekar S, Tsai J, Austin RC, Adeli K** 2005 Glucosamine-induced endoplasmic reticulum stress promotes ApoB100 degradation: evidence for Grp78-mediated targeting to proteasomal degradation. Arterioscl Throm Vas 25:571–577
41. **Gregor MF, Yang L, Fabbrini E, Mohammed BS, Eagon JC, Hotamisligil GS, Klein S** 2009 Endoplasmic reticulum stress is reduced in tissues of obese subjects after weight loss. Diabetes 58:693–700
42. **Werstuck GH, Lentz SR, Dayal S, Hossain GS, Sood SK, Shi YY, Zhou J, Maeda N, Krisans SK, Malinow MR, Austin RC** 2001 Homocysteine-induced endoplasmic reticulum stress causes dysregulation of the cholesterol and triglyceride biosynthetic pathways. J Clin Invest 107:1263–1273
43. **Kim R, Emi M, Tanabe K, Murakami S** 2006 Role of the unfolded protein response in cell death. Apoptosis 11:5–13
44. **Rao RV, Hermel E, Castro-Obregon S, del Rio G, Ellerby LM, Ellerby HM, Bredesen DE** 2001 Coupling endoplasmic reticulum stress to the cell death program. Mechanism of caspase activation. J Biol Chem 276:33869–33874
45. **Rao RV, Poksay KS, Castro-Obregon S, Schilling B, Row RH, del Rio G, Gibson BW, Ellerby HM, Bredesen DE** 2004 Molecular components of a cell death pathway activated by endoplasmic reticulum stress. J Biol Chem 279:177–187
46. **Laybutt DR, Preston AM, Akerfeldt MC, Kench JG, Busch AK, Biankin AV, Biden TJ** 2007 Endoplasmic reticulum stress contributes to β cell apoptosis in type 2 diabetes. Diabetologia 50:752–763
47. **Okada K, Minamino T, Tsukamoto Y, Liao Y, Tsukamoto O, Takashima S, Hirata A, Fujita M, Nagamachi Y, Nakatani T, Yutani C, Ozawa K, Ogawa S, Tomoike H, Hori M, Kitakaze M** 2004 Prolonged endoplasmic reticulum stress in hypertrophic and failing heart after aortic constriction: possible contribution of endoplasmic reticulum stress to cardiac myocyte apoptosis. Circulation 110:705–712
48. **Winnay JN, Boucher J, Mori MA, Ueki K, Kahn CR** 2010 A regulatory subunit of phosphoinositide 3-kinase increases the nuclear accumulation of X-box-binding protein-1 to modulate the unfolded protein response. Nat Med 16:438–445
49. **Ron D, Walter P** 2007 Signal integration in the endoplasmic reticulum unfolded protein response. Nat Rev Mol Cell Biol 8:519–529
50. **Ozcan U, Cao Q, Yilmaz E, Lee AH, Iwakoshi NN, Ozdelen E, Tuncman G, Görgün C, Glimcher LH, Hotamisligil GS** 2004 Endoplasmic reticulum stress links obesity, insulin action, and type 2 diabetes. Science 306:457–461
51. **Schenk S, Saberi M, Olefsky JM** 2008 Insulin sensitivity: modulation by nutrients and inflammation. J Clin Invest 118:2992–3002
52. **Lin J, Yang R, Tarr PT, Wu PH, Handschin C, Li S, Yang W, Pei L, Uldry M, Tontonoz P, Newgard CB, Spiegelman BM** 2005 Hyperlipidemic effects of dietary saturated fats mediated through PGC-1β coactivation of SREBP. Cell 120:261–273

53. **Sinha S, Perdomo G, Brown NF, O'Doherty RM** 2004 Fatty acid-induced insulin resistance in L6 myotubes is prevented by inhibition of activation and nuclear localization of nuclear factor κB. J Biol Chem 279:41294–41301
54. **Wei Y, Wang D, Topczewski F, Pagliassotti MJ** 2006 Saturated fatty acids induce endoplasmic reticulum stress and apoptosis independently of ceramide in liver cells. Am J Physiol 291:E275–E281
55. **Wang D, Wei Y, Pagliassotti MJ** 2006 Saturated fatty acids promote endoplasmic reticulum stress and liver injury in rats with hepatic steatosis. Endocrinology 147:943–951
56. **Thastrup O, Cullen PJ, Drøbak BK, Hanley MR, Dawson AP** 1990 Thapsigargin, a tumor promoter, discharges intracellular Ca2+ stores by specific inhibition of the endoplasmic reticulum Ca2(+)-ATPase. Proc Natl Acad Sci USA 87:2466–2470
57. **Ye R, Jung DY, Jun JY, Li J, Luo S, Ko HJ, Kim JK, Lee AS** 2010 Grp78 heterozygosity promotes adaptive unfolded protein response and attenuates diet-induced obesity and insulin resistance. Diabetes 59:6–16
58. **Nakatani Y, Kaneto H, Kawamori D, Yoshiuchi K, Hatazaki M, Matsuoka TA, Ozawa K, Ogawa S, Hori M, Yamasaki Y, Matsuhisa M** 2005 Involvement of endoplasmic reticulum stress in insulin resistance and diabetes. J Biol Chem 280:847–851
59. **Morand JP, Macri J, Adeli K** 2005 Proteomic profiling of hepatic endoplasmic reticulum-associated proteins in an animal model of insulin resistance and metabolic dyslipidemia. J Biol Chem 280: 17626–17633
60. **Zhang L, Perdomo G, Kim DH, Qu S, Ringquist S, Trucco M, Dong HH** 2008 Proteomic analysis of fructose-induced fatty liver in hamsters. Metabolism 57:1115–1124
61. **Bertolotti A, Zhang Y, Hendershot LM, Harding HP, Ron D** 2000 Dynamic interaction of BiP and ER stress transducers in the unfolded-protein response. Nat Cell Biol 2:326–332
62. **Oyadomari S, Yun C, Fisher EA, Kreglinger N, Kreibich G, Oyadomari M, Harding HP, Goodman AG, Harant H, Garrison JL, Taunton J, Katze MG, Ron D** 2006 Cotranslocational degradation protects the stressed endoplasmic reticulum from protein overload. Cell 126:727–739
63. **Merksamer PI, Trusina A, Papa FR** 2008 Real-time redox measurements during endoplasmic reticulum stress reveal interlinked protein folding functions. Cell 135:933–947
64. **Luciani DS, Gwiazda KS, Yang TL, Kalynyak TB, Bychkivska Y, Frey MH, Jeffrey KD, Sampaio AV, Underhill TM, Johnson JD** 2009 Roles of IP3R and RyR Ca2+ channels in endoplasmic reticulum stress and β-cell death. Diabetes 58:422–432
65. **Sage AT, Walter LA, Shi Y, Khan MI, Kaneto H, Capretta A, Werstuck GH** 2010 Hexosamine biosynthesis pathway flux promotes endoplasmic reticulum stress, lipid accumulation, and inflammatory gene expression in hepatic cells. Am J Physiol 298:E499–E511
66. **Song B, Scheuner D, Ron D, Pennathur S, Kaufman RJ** 2008 Chop deletion reduces oxidative stress, improves β cell function, and promotes cell survival in multiple mouse models of diabetes. J Clin Invest 118:3378–3389
67. **Hotamisligil GS** 2007 Endoplasmic reticulum stress and inflammation in obesity and type 2 diabetes. Novartis Found Symp 286:86–94
68. **Scheuner D, Kaufman RJ** 2008 The unfolded protein response: a pathway that links insulin demand with β-cell failure and diabetes. Endocr Rev 29:317–333